ARGUMENTIEREN BEIM EXPERIMENTIEREN IN DER PHYSIK

—

Die Bedeutung personaler und situationaler Faktoren

DISSERTATION

zur Erlangung des akademischen Grades

doctor rerum naturalium
(Dr. rer. nat.)

im Fach Physik

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

von

TOBIAS LUDWIG

Präsidentin der Humboldt-Universität zu Berlin
Prof. Dr.-Ing. Dr. Sabine Kunst

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät
Prof. Dr. Elmar Kulke

Gutachter:
1. Prof. Dr. Burkhard Priemer
2. Prof. Dr. Doris Lewalter
3. Prof. Dr. Alexander Kauertz

Tag der mündlichen Prüfung: 12. Juli 2017

ZUSAMMENFASSUNG

Argumentieren ist zentraler Bestandteil naturwissenschaftlicher Erkenntnisgewinnung. Dennoch gibt es bisher nur wenige Forschungsarbeiten, die untersuchen, wie Lernende auf der Grundlage selbstständig durchgeführter Experimente für bzw. gegen eine eigene Hypothese argumentieren.

Vor diesem Hintergrund untersucht diese Arbeit anhand eines physikalischen Experiments zum Fadenpendel, welchen Einfluss personale Faktoren und die Art der Lernumgebung darauf nehmen, welche Typen von Argumenten verwendet werden. Die in dieser Arbeit untersuchten personalen Faktoren sind das Fachwissen, das situationale Interesse, das Kognitionsbedürfnis und die Werteinschätzung der Naturwissenschaften. Bei den Lernumgebungen werden reale und virtuelle Experimente unterschieden. Die auf ihre Verwendung hin untersuchten Argumente fallen in die vier Kategorien „Intuition“, „Expertenwissen“, „Messunsicherheiten (explizit)“ sowie „Daten als Evidenz“. Ferner wird in dieser Arbeit untersucht, inwiefern die Verwendung dieser Argumentkategorien den Lernerfolg beeinflusst.

Zur Beantwortung der aufgeführten Forschungsfragen dokumentiert diese Arbeit drei aufeinander aufbauende Studien. Auf der Basis von Interviewdaten konnten zunächst für die von Schülerinnen und Schüler vorgebrachten Argumente beim Wechseln bzw. Beibehalten eigener Hypothesen beim Experimentieren zehn Kategorien identifiziert werden. Zur quantitativen Erfassung wurde dann für die vier o. g. Argumentkategorien ein Likert-skaliertes Instrument entwickelt. Die aufgeführten Fragestellungen wurden schließlich in einer randomisierten Studie mit 938 Schülerinnen und Schülern untersucht.

Bei der Untersuchung des Einflusses personaler Faktoren zeigt sich, dass Schülerinnen und Schüler in einer Argumentation für bzw. gegen eine physikalische Hypothese umso eher Daten als Evidenz heranziehen, je höher das fachliche Vorwissen ist. Die Verwendung dieser Argumentkategorie erhöht wiederum die Wahrscheinlichkeit dafür, dass Lernende nach dem Experimentieren eine fachlich adäquate Hypothese aufstellen. Dies impliziert, dass der Umgang mit experimentellen Daten und Beobachtungen im Physikunterricht stärker als bisher berücksichtigt werden muss, z. B. durch eine explizitere Förderung von Fähigkeiten zum Umgang mit experimentellen Daten. Bis auf einen gut erklärbaren Unterschied können grundlegende Unterschiede beim Experimentieren zwischen Gruppen, die mit einem Real- bzw. Computerexperiment gearbeitet haben, nicht belegt werden.

Die Studie trägt zu einem besseren Verständnis des Argumentierens beim Experimentieren und den damit verbundenen epistemischen Prozessen bei.

ABSTRACT

Argumentation from data and evidence evaluation is widely seen as a scientific core practice. One approach to engage students in a meaningful argumentation practice is to provide lab work situations where students can construct hypotheses on the basis of their own prior knowledge and consequently evaluate these hypotheses in light of self-collected data and experimental observations. However, until recently, only little research has analyzed students' argumentation from data.

Against this backdrop this research seeks to identify: a) the influences of personal factors (such as content knowledge, need for cognition, situational interest, and personal relevance); b) the type of learning environment as a situational factor (real vs. virtual experiment) on the use of different categories of argument (such as Intuition, Appeal to Authority, Measurement Uncertainties (explicit) and Data as Evidence); c) the influence of argumentation on learning outcomes through experimentation in school labs.

To answer these questions, this thesis reports on a set of subsequently conducted studies where students conducted a physical experiment. First, an interview-study was used to identify the different types of arguments used by students. Analyses focused on the nature of justification in argument (Sampson & Clark, 2008) and revealed ten different categories students use while arguing for or against hypotheses. As a next step, four out of ten categories were operationalized by means of a Likert-scaled instrument to assess the use of different types of argument in a valid and reliable manner. The findings from a randomized study among 938 secondary school students in a lab work setting indicate, among others, that content knowledge is positively related to the use of data as evidence. Again, the use of data as evidence increases the probability of stating a correct hypothesis after conducting the experiment. This implies that the ability to deal with data and measurement uncertainties should be better fostered in physics classes. Besides one explicable difference, no evidence was found, which supports the hypothesis that students' argumentation would differ while working with hands-on materials vs. computer simulations.

This study contributes to a better understanding of argumentation from data in school labs and the learning of science through experimentation.

INHALTSVERZEICHNIS

ABBILDUNGSVERZEICHNIS

TABELLENVERZEICHNIS

ABKÜRZUNGEN

Die Abkürzungen von latenten Variablen sind klein gedruckt.

ANOVA Varianzanalyse

ANCOVA Kovarianzanalyse

BIC Bayesian Information Criterion

CEST Cognitive-Experiential-Self-Theorie

CFA Konfirmatorische Faktorenanalyse

CFI Comparative-Fit-Index

CRKM Cognitive Reconstruction of Knowledge Model

ELM Elaboration-Likelihood Model of Persuasion

evi Argumentkategorie Daten als Evidenz

exp Argumentkategorie Expertenwissen

FIML Full-Information-Maximum-Likelihood

fw Fachwissen Mechanik

HSM Heuristisch-Systematisches Modell

int Argumentkategorie Intuition

IRT Item-Response-Theorie

LL Log-Likelihood

MANOVA Multivariate Varianzanalyse

MANCOVA Multivariate Kovarianzanalyse

MAR Missing-at-Random

MG-SGM Multigruppen-Strukturgleichungsmodelle

MG-CFA Multigruppen-konfirmatorische Faktorenanalyse

ML Maximum-Likelihood

MML Marginal-Maximum-Likelihood

MNSQ Mean-Square

mu Argumentkategorie Messunsicherheiten (explizit)

nfc Kognitionsbedürfnis

RMSEA Root Mean Square Error of Approximation

SDDS Scientific-Discovery-as-Dual-Search

si Situationales Interesse

SGM Strukturgleichungsmodelle

SRMR Standardized Root Mean Square Residual

wdn Werteinschätzung der Naturwissenschaften

WLE Weighted-Likelihood-Estimator

WLS Weighted-Least-Squares

1 EINLEITUNG

I think this has to be the slowest "Aha!" in the history of science, in that nowadays, when you're looking at data, it's a very complex task to interpret it very carefully. (...)And the first job we had when we saw the data coming in was to say "Ah, well, it looks like good data, and this should be nice, but right now it doesn't make any sense." (...) It's a very different result than anything we ever expected. (...) But we also felt very strongly the tension between wanting to announce an important answer, but also wanting to make sure that whatever you announced, people would believe it and that you would believe it, and that you had done all the homework necessary to make it make sense.

— Saul Perlmutter, Physik-Nobelpreisträger 2011

Mit diesen Worten kommentierte Saul Perlmutter seine Nominierung für den Physik-Nobelpreis für die Entdeckung der beschleunigten Expansion des Universums im Jahr 2011 (Nobel Media AB, 2014). Perlmutter wollte anhand neu gewonnener Daten des Hubble-Weltraumteleskops einen Beleg für die rund 70 Jahre alte Annahme finden, dass die durch den Urknall bedingte Ausdehnung des Universums aufgrund der Gravitationskraft langsamer wird und irgendwann zum Erliegen kommt. Was Perlmutter und Forschergruppen um seine Kollegen Schmidt und Riess in den Daten allerdings fanden, war kein Beleg für diese Annahme, sondern für das Gegenteil: eine beschleunigte Expansion des Universums. In der Folge musste Perlmutter nicht nur sich und sein Team auf Grundlage der Daten von der neuen Hypothese überzeugen, sondern auch die wissenschaftliche Community.

Das hier angeführte Beispiel aus jüngster Zeit illustriert einen bedeutsamen epistemischen Prozess in den Naturwissenschaften: Werden Daten in Verbindung mit Hypothesen gebracht, indem sie eine Hypothese stützen oder widerlegen, handelt es sich dabei um die Entwicklung eines *wissenschaftlichen Arguments* (Osborne, 2010; Toulmin, 1958).

Es wird seit längerer Zeit gefordert, das naturwissenschaftliche Argumentieren stärker im Unterricht und in der naturwissenschaftsdidaktischen Forschung zu berücksichtigen (Driver, Newton & Osborne, 2000; Kuhn, 1993). Dies geschieht zum einen aufgrund der Bedeutung des Argumentierens als *scientific practice* (Berland & Reiser, 2011; Kelly, 2008). Begründet wird dies mit der Annahme, dass naturwissenschaftliches Argumentieren als epistemischer Prozess nicht nur das konzeptuelle Verständnis fördert (Ford, 2008; Weinberger & Fischer, 2006), sondern Lernenden auch die Möglichkeit bietet, über die Natur der Naturwissenschaften zu lernen (Manz, 2014): So wird

z. B. davon ausgegangen, dass das oftmals vorherrschende positivistische Bild der Naturwissenschaften als "set of facts and conclusions" (McComas, 2004, S. 26) überwunden und ersetzt werden kann durch eines, das dem vorläufigen und diskursiven Charakter naturwissenschaftlicher Erkenntnis eher gerecht wird (Driver et al., 2000); Lemke (1990, S. 134) fasst naturwissenschaftliche Erkenntnis treffend folgendermaßen zusammen: "It is fallible, often uncertain, and sometimes creatively ambiguous."

Zum anderen ist es ein Ziel einer *scientific literacy*, dass Schüler befähigt werden, Ansichten und Aussagen zu naturwissenschaftlichen Themen evidenzbasiert begründen zu können (Bybee, 1997; DeBoer, 2000; Schiepe-Tiska, Schöps, Rönnebeck, Köller & Prenzel, 2012) – ein Aspekt, der durch das naturwissenschaftliche Argumentieren ebenfalls adressiert wird.

Aufgrund des hohen didaktischen Potenzials rückte das naturwissenschaftliche Argumentieren verstärkt in den Fokus der fachdidaktischen Forschung. Dabei wurden verschiedene Fragestellungen untersucht, z. B. zu den Möglichkeiten der Analyse von Argumentationen (z. B. Kelly, Regev & Prothero, 2007; Sampson & Clark, 2008), zu der Förderung argumentativer Fähigkeiten (z. B. Iordanou & Constantinou, 2015; Osborne, 2013; Ryu & Sandoval, 2012; Tröbst, Hardy & Möller, 2011; Zohar & Nemet, 2002), hinsichtlich der Argumentationsfähigkeit beim Schreiben (z. B. Heitmann, Hecht, Schwanewedel & Schipolowski, 2014; Kelly & Takao, 2002) oder die Entwicklung argumentativer Fähigkeiten (z. B. Koslowski, 1996; Kuhn & Udell, 2003; Osborne, Donovan, Henderson, MacPherson & Wild, 2016; Zimmerman, 2007).

Es wird jedoch kritisiert, dass das angeführte Ziel im Hinblick auf die Vermittlung eines adäquaten Bildes über Naturwissenschaften gar nicht erreicht werden könne, da das Argumentieren in Lehr-Lern-Situationen nicht mit dem *naturwissenschaftlichen* Argumentieren zu vergleichen sei: "We think it unlikely that people who do not practice science are likely to engage in truly scientific argumentation" (Sandoval & Millwood, 2007, S. 71). Diese Kritik wird insbesondere aus einer soziokulturellen Perspektive auf das Lernen geäußert (Lave & Wenger, 1991; Wertsch, 1991), in der berücksichtigt wird, dass sich Lernprozesse in der Schule und wissenschaftliche Aktivitäten im Hinblick auf die Normen und Ziele unterscheiden (Chinn, Buckland & Samarapungavan, 2011; Sandoval, 2005) und dementsprechend kein „authentischer" wissenschaftlicher Erkenntnisprozess zu erwarten ist (Abd-El-Khalick, 2008; Manz, 2014).

Lernsituationen, in denen das naturwissenschaftliche Experiment als Argumentationsgelegenheit eingesetzt wird, stellen eine Möglichkeit dar, dieser Kritik zu begegnen. In solchen Fällen werden wesentliche Elemente eines Arguments, nämlich Daten und Hypothesen, quasi „in situ" erzeugt. Lernende konstruieren auf der Basis von

Vorerfahrungen Hypothesen, überprüfen diese in einem naturwissenschaftlichen Experiment und argumentieren im Anschluss für oder gegen diese zuvor aufgestellten Hypothesen. Dennoch gibt es bisher nur wenige Forschungsarbeiten, die explizit das Argumentieren beim Experimentieren untersuchen: Die Qualität von Argumentationen untersuchen Kind, Kind, Hofstein und Wilson (2011) anhand struktureller Analysen, den Einfluss des Grades der Offenheit der Aufgabenstellung beim Experimentieren auf das Argumentieren untersuchen Katchevich, Hofstein und Mamlok-Naaman (2013).

Es ist daher bisher weitgehend unbekannt, a) welche Typen von Argumenten Lernende auf der Basis experimenteller Daten und Beobachtungen überhaupt generieren, b) wie diese Argumente von personalen bzw. situationalen Faktoren beeinflusst werden und c) welchen Einfluss diese Argumente auf den Lernerfolg, z. B. in Form der Richtigkeit der nach dem Experiment aufgestellten Hypothese, nehmen. Diese Fragestellungen werden durch die vorliegende physikdidaktische Arbeit adressiert. Die Kenntnis darüber, welche personalen Faktoren (wie das Fachwissen oder das situationale Interesse) bzw. welche situationalen Faktoren (z. B. das Experimentiermedium – reales Experiment vs. Computersimulation) Einfluss auf die Art der Argumentation (z. B. das Heranziehen von Daten als Evidenz oder die Begründung auf Grundlage von Intuition) nehmen, ist aus fachdidaktischer und schulpraktischer Perspektive hochrelevant. Entsprechend fordern Sampson und Clark (2008, S. 468): "we need to better understand the criteria that students use to determine what evidence is most persuasive or to warrant one idea over another." Auf diese Weise lässt sich identifizieren, wie Lernarragenements konstruiert werden können, damit zum einen die im Konzept der *scientific literacy* genannten Ziele hinsichtlich des evidenzbasierten naturwissenschaftlichen Argumentierens erreicht werden und zum anderen ein adäquates Bild über die Natur der Naturwissenschaften vermittelt wird.

Zur Beantwortung der aufgeführten Fragestellungen wurden in der vorliegenden Arbeit 938 Schülerinnen und Schüler der Mittelstufe bei der Durchführung eines einfachen physikalischen Experiments (Zusammenhang zwischen Pendelmasse und Schwingungsdauer) beobachtet und befragt. Der Forschungsgegenstand, nämlich das Argumentieren für oder gegen eine zuvor aufgestellte Hypothese auf der Grundlage von Informationen, die aus dem Experiment gewonnen wurden, ergibt sich erst durch die Interaktion von Individuen mit der Experimentiersituation. Die vorliegende Arbeit nimmt daher eine situierte Perspektive ein. Damit wird der Tatsache Rechnung getragen, dass Lernen von Naturwissenschaften – hier durch Experimentieren – nicht von der Umwelt und Situation, in der es stattfindet, abstrahiert werden kann (Chinn et al., 2011; Greeno, 1998; Lave & Wenger, 1991; Sadler, 2009).

ANMERKUNG ZU SCHREIBWEISEN UND NOTATION: Im Folgenden wird in der vorliegenden Arbeit ausschließlich die grammatisch männliche Form verwendet. Sie soll explizit als genderunabhängig verstanden werden.

Die Präsentation der Ergebnisse orientiert sich an der in Naturwissenschaftsdidaktik und Lehr-Lern-Forschung üblichen anglo-amerikanische Schreibweise. Es wird daher – auch wenn es im Deutschen unüblich ist – als Dezimaltrennzeichen ein Punkt verwendet. Zudem wird für Werte statistischer Größen, die definitionsgemäß nur zwischen -1 und 1 liegen können, die führende Null nicht ausgeschrieben.

2

THEORETISCHER HINTERGRUND UND STAND DER FORSCHUNG

Bei der Darstellung des theoretischen Hintergrunds und des aktuellen Forschungsstandes werden Schwerpunkte auf die zum Verständnis der vorliegenden Arbeit relevanten Konzeptionen gelegt. Eher kurz ist daher der allgemeine Teil zum Argumentieren. Ausführlicher werden hingegen Arbeiten zum Argumentieren in Situationen mit Experimenten dargestellt.

Die Beschreibung der theoretischen Grundlagen gliedert sich in die folgenden Bereiche: In Abschnitt 2.1 wird zunächst auf die hier verwendete theoretische Konzeption des Argumentierens eingegangen, dann wird die Befundlage zum Argumentieren beim Experimentieren dargelegt. In der vorliegenden Arbeit wird u. a. untersucht, inwiefern das Argumentieren durch den situationalen Faktor Experimentiermedium beeinflusst wird. Abschnitt 2.2 stellt die Motivation für dieses Vorgehen dar. Abschnitt 2.3 beschreibt die Vorarbeiten zur Kategorisierung der beim Experimentieren vorgebrachten Argumente. Auf dieser Grundlage wird in Abschnitt 2.4 und Abschnitt 2.5 die Übertragung des Elaboration-Likelihood Model of Persuasion (ELM) auf das Argumentieren beim Experimentieren beschrieben. In Abschnitt 2.6 werden Theorie und Vorarbeiten zusammengeführt und das Forschungsinteresse abgeleitet.

2.1 DAS NATURWISSENSCHAFTLICHE ARGUMENTIEREN

2.1.1 *Definition des Begriffs*

Es existieren eine Reihe unterschiedlicher theoretischer Konzeptionen zum Argumentieren. Dies ist insbesondere der Tatsache geschuldet, dass das Argumentieren in ganz unterschiedlichen Forschungsrichtungen betrachtet wird, z. B. in der Philosophie, der Lehr-Lernforschung oder den Didaktiken der Naturwissenschaften. Unter dem Begriff „Argumentation“ wird dabei meist der *Prozess* verstanden, bei dem Argumente gebildet werden. Unter einem Argument wiederum versteht man das *Produkt* eines argumentativen Prozesses, welcher eine Behauptung („claim“) auf Grundlage von Evidenzen stützt (Osborne & Patterson, 2011; Sampson & Clark, 2008; Toulmin, 1958). Diese Evidenzen können auf der Grundlage empirischer Daten und Schlussregeln („warrants“, hier: Begründungen) gebildet werden. Ein Argument ist daher eine Aussage, welche versucht, durch das In-Beziehung-setzen der genannten Elemente eine Behauptung zu be-

gründen (Prechtl, 2016). Ein Kriterium, welches das Argumentieren von anderen epistemischen Instanzen wie z. B. dem Erklären abgrenzt, ist die zugrundeliegende Absicht („epistemic goal"): Während beim Erklären angestrebt wird, ein tieferes Verständnis eines beobachteten (naturwissenschaftlichen) Phänomens zu erlangen oder dieses sachlogisch inhaltlich herzuleiten, wird beim Argumentieren versucht, in einer Überzeugungsabsicht eine Behauptung zu rechtfertigen (Osborne & Patterson, 2011). Oft liegt dabei eine strittige Situation vor. Zur Erzeugung von strittigen Situationen werden in der vorliegenden Arbeit Experimente mit nicht-hypothesenkonformen Ergebnissen („anomalous data", Chinn & Brewer, 1998) verwendet (eine tiefergehende Darstellung dazu folgt in Abschnitt 2.1.5).

Es sei hier insbesondere auf die Ausführungen von Gromadecki (2009) zur weiteren Abgrenzung des Argumentationsbegriffs von anderen Begriffen wie dem Begründen, Erklären, Beweisen und Rechtfertigen verwiesen. Tiefergehende Darstellungen aus philosophischer Perspektive finden sich z. B. bei Benoit, Hample und Benoit (1992), Kuhn und Udell (2003), van Eemeren und Grootendorst (2004), Walton (1990, 2008, 2016).

2.1.2 *Interpersonelle vs. intrapersonelle Persuasion als Ziel von Argumentation*

In den Naturwissenschaften verfolgt der Prozess des Argumentierens das Ziel, sich oder die wissenschaftliche Community von einer bestimmten Hypothese zu überzeugen (Walton, 1990). Dabei ist das Argumentieren oftmals in soziale Prozesse eingebunden (Duschl, 2007; Kolstø & Ratcliffe, 2007). Beispielsweise argumentieren Naturwissenschaftler mit dem Ziel, andere Kollegen von einer bestimmten Erkenntnis zu überzeugen (Ryu & Sandoval, 2012), wie sich auch an dem einleitenden Zitat des Physik-Nobelpreisträgers zeigt (vgl. Kapitel 1). Daher unterstellen einige Definitionen (z. B. van Eemeren & Grootendorst, 2004) und Forschungsarbeiten (z. B. Riemeier, von Aufschnaiter, Fleischhauer & Rogge, 2012) dem Prozess des Argumentierens stets eine kommunikative, dialogische Komponente auf einer inter-individuellen Ebene. Das Argumentieren kann aber auch als eine innere Auseinandersetzung ohne Einbettung in Kommunikationsprozesse oder Dialogizität betrachtet werden, wie Schwarz und Asterhan (2010) formulieren (siehe auch Driver et al., 2000; Ford, 2012; Garcia-Mila & Andersen, 2007; Jiménez-Aleixandre & Erduran, 2007):

> Internal intra-personal activities of argumentation are, for example, conducting a virtual discussion between two sides in one's mind or when an individual declaratively and consciously weighs the reasons for and against a certain line of action, standpoint or solution. (S. 144)

In der vorliegenden Arbeit führen Lernende einzeln und ohne Einbindung in einen sozialen Kontext Experimente durch und argumentieren auf Grundlage von aus dem Experiment gewonnenen Informationen für oder gegen eine zuvor aufgestellte Hypothese. Es werden hier also Argumentationsprozesse auf der dargestellten intrapersonellen Ebene untersucht. Es sind zum derzeitigen Zeitpunkt keine Arbeiten bekannt, die das Argumentieren auf der intrapersonellen Ebene untersucht haben.

2.1.3 *Die 'nature of justification'-Perspektive*

Argumentation verfolgt nicht nur das Ziel der inter- bzw. intrapersonellen Persuasion, sondern dient auch dem Zweck, naturwissenschaftliche Erkenntnis zu *begründen*. Das Begründen spielt in den Naturwissenschaften eine zentrale Rolle, wie Jiménez-Aleixandre und Erduran (2007) betonen:

> In science, knowledge construction is linked to knowledge justification, and claims should be related either to a path of logical clauses or to data and evidence from different sources (or to both). Hence, argumentation in scientific topics can be defined as the connection between claims and data through justification or the evaluation of knowledge claims in light of evidence, either empirical or theoretical. (S.13)

Begründungen können daher als Elemente eines Arguments gesehen werden, die erklären, *warum* Daten oder Beobachtungen als Evidenz für eine Hypothese gelten können (McNeill & Krajcik, 2007) und bieten eine gute Möglichkeit, die Qualität von Argumenten zu analysieren (Ryu & Sandoval, 2012). Durch die Analyse der Art der Begründung wird ein Fokus auf die epistemische Dimension (Weinberger & Fischer, 2006) gelegt, was ermöglicht, im Gegensatz zu Arbeiten, die lediglich strukturelle Elemente im Toulmin-Schema identifizieren (Erduran, Simon & Osborne, 2004; Riemeier et al., 2012), auch zu analysieren, *wie* Lernende Wissen durch Argumentieren konstruieren.

Das Begründen spiegelt sich auch in den unterschiedlichen Ansätzen zur Analyse von Argumentationen wider. Sampson und Clark (2008) unterscheiden dabei 1. einen strukturellen Zugang, bei dem unter Rückgriff auf bestimmte Schemata (wie das von Toulmin, 1958, oder von Schwarz, Neuman, Gil und Ilya, 2003) die einzelnen Komponenten von Argumenten untersucht werden, 2. einen inhaltlichen Zugang, bei dem die fachliche Richtigkeit oder Adäquatheit aus naturwissenschaftlicher Perspektive untersucht wird, sowie 3. die Analyse der Natur der Begründungen, d. h. die Frage danach, wie Behauptungen innerhalb eines Arguments gestützt werden („nature of justification of a claim“, Sampson & Clark, 2008, S. 449). Insbesondere beim Lernen von Naturwissenschaften spielt das Begründen eine

zentrale Rolle, da es eher auf das konzeptuelle Verständnis als auf das Erinnern von Fakten abzielt (Weinberger & Fischer, 2006). Dies spiegelt sich auch in den aktuellen nationalen und internationalen Bildungsstandards wider (DE: Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, 2004, UK: Department for Education, 2015, USA: NGSS Lead States, 2013).

Diese Perspektive auf das Argumentieren unterliegt aber auch Beschränkungen. In der vorliegenden Arbeit werden Begründungen in Lehr-Lern-Situationen untersucht. Hier können andere Kriterien zur Analyse der Qualität einer Begründung gelten, denn ein „gutes" Argument bzw. eine „gute" Begründung im schulischen Kontext muss nicht zwingend auch ein „gute" Begründung in der Wissenschaft darstellen (Sandoval, 2003). Schulischer Unterricht folgt zudem oftmals einem induktiven Erkenntnisweg, der eher auf das Generieren von Hypothesen und Theorien als auf das Testen abzielt. Dies kann Lernende in der Art der gegebenen Begründungen beeinflussen. Weiterhin ist bekannt, dass Begründungen disziplinabhängig (Toulmin, 2003), kulturabhängig (Ryu & Sandoval, 2012) sowie situationsabhängig (Chinn et al., 2011) unterschiedlich ausfallen können. So ist z. B. bekannt, dass Lernende unterschiedliche Begründungen liefern können, wenn die epistemische Zielsetzung, unter der gearbeitet wird, sich ändert („epistemic goal", Chinn et al., 2011). Es zudem davon auszugehen, dass Begründungen auch bekannten *biases* beim Umgang mit Hypothesen unterliegen, z. B. dem *confirmation bias*, bei dem Lernende tendenziös vorgehen und eigene Hypothesen bestätigen möchten, oder dem *fear of rejection bias*, bei dem Lernende das Ablehnen eigener Hypothesen vermeiden (Ganser & Hammann, 2009; Hammann, Phan, Ehmer & Bayrhuber, 2006; Klayman & Ha, 1987; Njoo & de Jong, 1993; van Joolingen & de Jong, 1997).

Das Begründen in naturwissenschaftlichen Lehr-Lern-Situationen wurde beforscht (Kelly, 2008; Sandoval & Millwood, 2005; Zohar & Nemet, 2002). Chinn et al. (2011, S. 154) fassen zusammen, dass dabei allerdings häufig epistemische Überzeugungen „at a relatively coarse grain size" in den Fokus genommen wurden, und fordert Ansätze, die den vielfältige Wirkungszusammenhängen beim Begründen Rechnung tragen, z. B. durch den Einsatz hochauflösender Instrumente.

Daher werden in der vorliegenden Arbeit die von Schülern vorgebrachten Argumente nach der *Art der Begründung* klassifiziert und deren Verwendung untersucht (Sampson & Clark, 2008). In Vorgriff zu den Vorarbeiten (vgl. Abschnitt 2.3), in denen verschiedene Typen der Begründung identifiziert werden konnten, soll dies an drei Beispielen deutlich gemacht werden: Eine Art der Begründung einer naturwissenschaftlichen Hypothese ist das Heranziehen von selbstständig generierten Daten als Evidenz. Weitere Arten der Begründung ist

der Verweis auf den Aufbau des Experiments oder die Begründung „aus dem Bauch heraus“, also ein Begründen auf Grundlage intuitiver Prozesse.

Im weiteren Verlauf dieser Arbeit wird der Oberbegriff „Argumentkategorie“ zur Unterscheidung verschiedener Typen von Begründungen verwendet.

2.1.4 *Das Experiment als Argumentationsgelegenheit*

In den bestehenden Modellen des Experimentierens in Lernsituationen (siehe z. B. Emden, 2011; Hammann, 2004, 2007, 2010; Hodson, 1993; Klahr & Dunbar, 1988; Lunetta, Hofstein & Clough, 2007; Mayer, Keiner, Ziemek & Klee, 2003; Muckenfuß, 1995; Nawrath, Maiseyenka & Schecker, 2011; Schmidkunz & Lindemann, 2003; Schreiber, Theyßen & Schecker, 2009; Tesch & Duit, 2004) werden der Umgang mit Hypothesen, die Evaluation von Messdaten und das Heranziehen der Messdaten als Evidenzen als grundlegende experimentelle Teiltätigkeiten genannt. Anmerkung: In der naturwissenschaftsdidaktischen Community wird das Experimentieren überwiegend aus dem Blickwinkel des kritischen Rationalismus im Popper'schen Sinne betrachtet (Chalmers, 2007). Dies zeigt sich z. B. in den erwähnten Modellen des Experimentierens, in denen das Testen von Hypothesen eine zentrale Rolle einnimmt. „Hypothesen zu testen kann dagegen nur als eine von mehreren experimentellen Strategien verstanden werden“ (Höttecke & Rieß, 2015, S. 136). Ein exploratives, nichttheoriegeleitetes Experimentieren ist aber aus epistemischer Perspektive ebenso legitim und in der Geschichte ebenfalls häufig anzutreffen (Lederman et al., 2014; Steinle, 1997). Die vorliegende Arbeit beschränkt das Experimentieren jedoch auf das hypothesenprüfende Vorgehen.

Beim Experimentieren werden wesentliche Elemente eines Arguments nach Toulmin (1958) „in situ“ produziert. Dies bezieht sich sowohl auf Daten als auch auf Hypothesen, die – zumindest im schulischen Kontext – auf der Basis von im Alltag etablierten Präkonzepten aufgestellt werden können. Das In-Beziehung-setzen von a priori bzw. a posteriori aufgestellten Hypothesen mit experimentellen Daten und Beobachtungen kann daher als Argumentationsprozess aufgefasst werden (Gott & Duggan, 2007). Dieser Aspekt bezieht sich auf den bereits in der Einleitung dargelegten Standpunkt, nach dem das Experiment im naturwissenschaftlichen Unterricht als eine gute Gelegenheit zum Argumentieren gelten kann – als ein *natural locus*, wie Kind et al. (2011, S. 2531) treffend konstatieren. Damit unterscheidet sich eine Lernsituation mit naturwissenschaftlichen Experimenten, in der Schüler eigenständig Hypothesen generieren und Messdaten aufnehmen, grundlegend von argumentativen Prozessen und Lernsituationen, in denen verschiedene Hypothesen bzw. Standpunkte sowie

Daten *vorgelegt* werden, wie dies in einer Vielzahl von Forschungsarbeiten zum Argumentieren geschieht (Chinn & Brewer, 1998; Sandoval & Millwood, 2005; Sandoval & Reiser, 2004). Die Übertragbarkeit der Ergebnisse dieser Studien auf das Argumentieren beim Experimentieren ist daher infrage zu stellen. Dies gilt insbesondere deshalb, weil eigene experimentelle Erfahrungen in Settings, die ohne eigenständige Durchführung eines Experiments auskommen, keine Rolle spielen können (Gott & Duggan, 2007):

> Again these exercises are of value in relation to argumentation and scientific literacy but they tend not to rely on primary data, thus making it difficult, if not impossible, to question the warrants, qualifiers and backings that lie behind the claims. (S. 287)

Die Bedeutung des Argumentierens beim Experimentieren ist in der fachdidaktischen Forschung erkannt worden. So benennen Lunetta et al. (2007, S. 402) „argumentation from data“ erstmals als ein Ziel des Experimentierens (vgl. Hofstein, 2017; Hofstein & Lunetta, 2004). Trotz des offenbar großen didaktischen Potentials wird das Experiment aber bisher kaum als Argumentationsgelegenheit genutzt: “scientific argumentation, which should be a natural part of any inquiry process has played a minor role in laboratory teaching” (Kind et al., 2011, S. 2548). Analog dazu berichten Watson, Swain und McRobbie (2004), dass argumentative Prozesse beim Experimentieren nahezu nicht auftreten. Auch gibt es bisher nur wenige Forschungsarbeiten, die explizit das Argumentieren beim Experimentieren untersuchen. Einige der wenigen Arbeiten zum Argumentieren beim Experimentieren, die eine eigenständige Durchführung des Experiments beinhaltet, ist die Arbeit von Kind et al. (2011). Dabei wurde die Qualität der Argumentation auf Grundlage des Toulmin-Schemas analysiert (vgl. Erduran et al., 2004). Variiert wurde die Stelle, an der eine Argumentation während des Experimentierens initiiert wurde. In einer ersten Gruppe trafen Lernende auf komplexe, mehrdeutige Daten (Temperaturabfall in mit einer warmen Flüssigkeit gefüllten Behältern unterschiedlicher Beschichtung). Die Idee war dabei, dass verschiedene Positionen eingenommen werden konnten, deren Standpunkt es argumentativ zu vertreten galt. In einer weiteren Bedingung wurde ein Experiment eingesetzt, das ein weit verbreitetes Präkonzept aufgreift (Verlust von Masse beim Lösen von Stoffen in Wasser). Dazu wurde durch das Aufstellen von Hypothesen im Vorfeld des Experiments eine argumentative Auseinandersetzung angeregt. In einer dritten Gruppe wurden Protokolle vorgelegt, die im gleichen Kontext wie in Gruppe 1 die Abhängigkeit des Temperaturabfalls eines mit warmen Wasser gefüllten Bechers von seiner äußeren Beschichtung untersuchen. Hier wurde die Auseinandersetzung ohne eigentliche Durchführung des Experiments angeregt, indem die Probanden die Validität dieser Daten aus zweiter Hand (Hug & Mc-

Neill, 2008; Pfeiler & Priemer, 2017) diskutieren sollten. Für die drei Situationen berichten Kind et al. (2011) von unterschiedliche Intensitäten argumentativer Auseinandersetzung. Die dritte Gruppe, die nicht selber experimentiert hatte, schnitt dabei am besten ab. Kind et al. (2011) schließen daraus, dass die eigenständige Durchführung eines Experiments das Argumentieren nicht zwingend fördert. Diese Aussage kann aber deswegen kritisiert werden, weil es sich bei den drei Gruppen um äußerst unterschiedlich gestaltete Szenarios handelte (Gruppe 3 hatte einen Zeitvorteil, da kein Experiment durchgeführt werden musste). Weiterhin berichten Kind et al. (2011), dass der argumentative Prozess an der Tatsache scheitert, dass es Studierenden schwer fällt, eigenständig von einem Prozess des naturwissenschaftlichen Experimentierens in einen anderen zu wechseln: Im Sinne des Scientific-Discovery-as-Dual-Search (SDDS)-Modells nach Klahr (2000) fällt der Wechsel von der Suche im Hypothesenraum zu einer Suche in den Experimentierraum und insbesondere die Kombination beider Prozesse zur Evaluation sehr schwer oder findet gar nicht statt. In der Konsequenz argumentieren Kind et al. (2011), dass argumentative Prozesse beim Experimentieren nur durch äußere Stimuli eintreten. Dies ist analog zu einer Studie von Kim und Song (2006), die in einer Studie zum Argumentieren bei offenen Experimentieraufgaben im Kontext der Chemie ebenfalls zu dem Schluss kommen, dass ein *scaffolding* zum Argumentieren beim Experimentieren nötig ist. Watson et al. (2004) vermuten als Ursache, dass Lernende das naturwissenschaftliche Experimentieren als starren und festgelegten Prozess sehen, der am Ende zu einer gesicherten Erkenntnis führt, und nicht als eine argumentative Auseinandersetzung. In Vorgriff auf Kapitel 4 sei hier erwähnt, dass u. a. auf Grundlage der hier referierten Ergebnisse ein Ansatz gewählt wurde, in dem eine Begründung für eine Hypothese durch einen äußeren Stimulus (in einem Interview bzw. Fragebogen) forciert wurde und nicht durch soziale Interaktion mit *peers*.

2.1.5 *Nicht-hypothesenkonforme Daten als Ausgangspunkt für Argumentationen*

Es gibt eine Reihe von Hinweisen zur Gestaltung von naturwissenschaftlichen Lernumgebungen unter konstruktivistischen Ansätzen, in denen sich Lernende durch Argumentation neue Inhalte erschließen (z. B. Jiménez-Aleixandre, 2007), sowie einige Werke mit konkreten Unterrichtsvorschlägen zum Argumentieren im naturwissenschaftlichen Unterricht (**Osborne2016**; Sampson, Enderle & Grooms, 2013). Zur Nutzung des naturwissenschaftlichen Experiments als Gelegenheit zum Argumentieren finden sich trotz der herausragenden Stellung des Experiments als Argumentationsgelegenheit bisher kaum Vorschläge.

Es stellt sich die Frage, welche Charakteristika von Experimenten überhaupt einen Anlass zum Argumentieren im naturwissenschaftlichen Unterricht bieten können. Eine Möglichkeit ist die Verwendung von Kontexten, zu denen Schüler bekanntermaßen fachinhaltlich defizitäre Präkonzepte in den Unterricht mitbringen. Widersprechen Daten zuvor aufgestellten Hypothesen, können diese als *nicht-hypothesenkonforme Daten* bezeichnet werden („anomalous data", Chinn & Brewer, 1993, 1998). Ein so entstehender kognitiver Konflikt kann ein Anlass für eine intensive argumentative Auseinandersetzung sein, der zum Ziel hat, diesen kognitiven Konflikt aufzulösen (z B. Kind et al., 2011; Walton, 1990). Den durch das Erleben von Anomalien hervorgerufenen kognitiven Konflikten wird nach Posner, Strike, Hewson und Gertzog (1982) ein starker Einfluss auf einen Konzeptwechsel unterstellt. Im Idealfall führt diese Auseinandersetzung dann zu einem Konzeptwechsel (Chinn & Brewer, 1998; Duit & Treagust, 2003; Guzzetti, Snyder, Glass & Gamas, 1993; Kanari & Millar, 2004). Nichtsdestotrotz hat ein einmal verankertes Konzept eine sehr hohe Persistenz, so dass ein Wechsel oftmals trotz auftretender Widersprüche ein schwerfälliger Prozess ist (Kang, Scharmann, Kang & Noh, 2010). Kanari und Millar (2004) konnten zeigen, dass Lernende eine unklare Datenlage zunächst nicht berücksichtigen und bei ihren Konzepten bleiben, die eher auf Erwartungen und dem Abrufen von Vorwissen beruhen. Dieses Vorgehen ist durchaus berechtigt und findet sich auch in der Naturwissenschaft (Chinn & Samarapungavan, 2001), wie auch an dem einleitenden Beispiel in Kapitel 1 deutlich wird. Erst ab einem gewissen Grad werden die Daten berücksichtigt, es kann ggf. zu einem Konzeptwechsel kommen („unambigous data", Chinn & Malhotra, 2002, S. 329).

Die Schwierigkeiten, die Lernende im Kontext des nicht-hypothesenkonformen Experimentierens zeigen, lassen sich zusätzlich auch auf den problematischen Umgang mit Hypothesen generell zurückführen (Murphy & Mason, 2006): Zum einen verhindert der *confirmation bias* eine allzu häufige Konfrontation mit unerwarteten Daten. Zum anderen entstammen aufgestellte Hypothesen oftmals „starken inhaltlichen Überzeugungen, die sich im täglichen Leben bewährt haben. Eine inadäquate Analyse experimenteller Daten wird in diesem Fall nicht durch mangelndes logisches Denken verursacht, sondern von dem Bestreben, bewährte Alltagskonzepte bestätigt zu finden" (Hammann et al., 2006, S. 297).

Die Reaktionen auf nicht-hypothesenkonforme Daten wurden in zahlreichen empirischen Studien untersucht (Chinn & Brewer, 1998; Lin, 2007; Mason, 2001). Chinn und Brewer (1998) haben 33 mögliche Reaktionen auf nicht-hypothesenkonforme Daten anhand von drei Fragen („Are the data accepted?", „Are the data explained?" und „Is the current theory changed?") einer Taxonomie zugeordnet. Diese Taxonomie ordnet die Reaktionen aufgrund der Vollständigkeit der

Annahme widersprüchlichen Theorie. Dabei ergaben sich die folgenden Stufen: *Ignoranz, Ablehnung, Unsicherheit, Ausschluss, Unentschiedenheit, Reinterpretation, schwacher Theoriewechsel, Theoriewechsel*. Lin (2007) ergänzt diese Taxonomie auf der Basis der gleichen Vorgehensweise um die Kategorie *Unsicherheit der Interpretation*. G. Lee und Byun (2011) ergänzen weiterhin die Kategorie *superficial theory change*. In diese Kategorie fallen Antworten, die erkennen lassen, dass ein Theoriewechsel zwar stattfindet, die Erklärungen dafür aber oberflächlich oder überhaupt nicht vorhanden sind.

Dennoch können bereits bekannte Systeme zur Klassifikation von Reaktionen auf nicht-hypothesenkonforme Daten nicht auf die vorliegende Arbeit übertragen werden, denn oft wurden diese Studien nicht im Kontext des (schulischen) naturwissenschaftlichen Experimentierens durchgeführt. Aus naturwissenschaftsdidaktischer Perspektive unterliegen diese Studien daher wesentlichen Einschränkungen: Oftmals wurde den Probanden zunächst ein umfangreicher Text vorgelegt, der den Lesern eine Eingangstheorie glaubhaft darstellen sollte (z. B. die Theorie, dass ein Meteoriteneinschlag das Aussterben der Dinosaurier verursacht hat). Im Verlauf der Untersuchung wurde den Probanden ein weiterer Text vorgelegt, der zur Eingangstheorie widersprüchliche Daten enthielt (z. B. fehlende Nachweise bei Bodenproben). Auf dieser Grundlage wurden die Probanden dann nach Begründungen für das Annehmen bzw. Ablehnen der anomalen Datenlage befragt. Dies trifft u. a. auf die Arbeiten von Chinn und Brewer (1998) und Mason (2001) zu. Problematisch an diesem methodischen Vorgehen ist, dass sich die Kombination aus nahegelegtem Präkonzept und vorgelegten nicht-hypothesenkonformen Daten z. B. nicht ohne Weiteres auf experimentelle Arbeitsweisen übertragen lässt. Im Gegensatz zu der bei „von außen“ erzeugten wissenschaftlichen Bewertungssituation sind im naturwissenschaftlichen Unterricht Präkonzepte bereits ohne explizite Thematisierung bei den Schülerinnen und Schülern fest verankert. Ein Wechsel erscheint hier schwieriger als bei einer erst kurzfristig zuvor angenommen Eingangshypothese. Zum anderen unterscheiden sich Fremddaten von selbstständig experimentell generierten Daten möglicherweise hinsichtlich ihrer Überzeugungskraft (Hug & McNeill, 2008). So lassen sich z. B. Zweifel an der Richtigkeit der Daten beim eigenständigen Experimentieren ggf. durch Wiederholen oder Änderungen im Versuchsaufbau ausräumen.

Lin (2007), Chinn und Malhotra (2002) und Shepardson (1999) hingegen untersuchten Reaktionen auf anomale Daten in einem Setting, bei dem Probanden experimentieren konnten. Diese Reaktionen wurden allerdings ebenfalls nach dem Schema von Chinn und Brewer (1998) klassifiziert, d. h. sie wurden im Hinblick auf den Vollzug eines Konzeptwechsels untersucht und nicht im Hinblick auf argumentative Prozesse. Berland und V. R. Lee (2010) untersuchten zwar ar-

gumentative Prozesse auf Grundlage nicht-hypothesenkonformer Daten, aber auch hier wurden Daten vorgelegt und konnten nicht selbstständig experimentell erhoben.

Bisher liegen daher keine Arbeiten vor, die *explizit* untersucht haben, welche Typen von Argumenten Lernende beim Experimentieren generieren. Daraus leitet sich die Relevanz der eigenen Vorarbeiten ab (Abschnitt 2.3), in denen in zwei physikalischen Experimentiersituationen Schüler befragt wurden, welche Argumente für oder gegen das Beibehalten einer zuvor aufgestellten Hypothese herangezogen werden.

2.2 DIE ROLLE VON COMPUTERSIMULATIONEN FÜR NATURWISSENSCHAFTLICHE LERNPROZESSE

In der vorliegenden Arbeit wird untersucht, welchen Einfluss die Art der Lernumgebung, d. h. ein reales Experiment bzw. eine Computersimulation, auf Argumentationen nimmt. Der folgende Abschnitt legt den theoretischen Rahmen für diese Fragestellung dar.

Experimentierprozesse werden in der Forschung zunehmend durch den Einsatz von computersimulierten Experimentierumgebungen untersucht. So wird in den Fachdidaktiken z. B. versucht, experimentelle Kompetenz durch das Arbeiten an Computersimulationen zu erfassen (z. B. Schecker, Neumann, Theyßen, Eickhorst & Dickmann, 2016; Schreiber, 2012; Schreiber et al., 2009). Die Lehr-Lern-Forschung untersucht dabei Aspekte wie die Strategieanwendung beim Experimentieren (Gößling, 2010; Künsting, Thillmann, Wirth, Fischer & Leutner, 2008), Zielspezifität beim Problemlösen (Künsting, Wirth & Paas, 2011; Wirth, Künsting & Leutner, 2009), das selbstregulierte Lernen durch Experimentieren (Thillmann, 2007), das *prompting* (Thillmann, Gößling, Wirth & Leutner, 2009) oder Fragestellungen bezüglich Lernprozessen beim entdeckenden Lernen (de Jong et al., 1999; de Jong & van Joolingen, 1998; Njoo & de Jong, 1993). Das Heranziehen von Computersimulationen als eine ökonomische Methode zur Untersuchung dieser Forschungsinteressen ist aus Sicht der Lehr-Lern-Forschung legitim. Einige Autoren dieser Studien verallgemeinern ihre Resultate vielfach explizit oder implizit auf reales Experimentieren, wie an den folgenden zwei Beispielen verdeutlicht werden soll. So schreiben van Joolingen und de Jong (1997, S. 307): "Discovery learning is easily facilitated in these simulation learning environments [...], because they allow the learner to actively engage in a scientific discovery process by doing experiments" . Unter „doing experiments" kann offensichtlich jedoch Unterschiedliches verstanden werden, wie sich auch bei Gößling (2010) zeigt (vgl. auch de Jong & van Joolingen, 1998, S. 180):

> Die Hauptaufgabe der Lernenden besteht beim Lernen mit den Lernumgebungen demnach darin, auf Charakte-

> ristika des der Simulation bzw. Lernumgebung zugrundeliegenden Modells zu schließen, in dem der Lernende die Werte der Inputvariablen verändert und die sich ändernden Outputwerte beobachtet [...]. Dadurch bietet sich den Lernenden die Möglichkeit, neues Wissen auf eine wissenschaftliche Weise zu generieren. (Gößling, 2010, S. 27)

Sowohl aus physikdidaktischer als auch aus erkenntnistheoretischer Perspektive kann aber angezweifelt werden, dass es sich bei dem hier geschilderten Vorgehen – der reinen Manipulation von Simulationen und dem Erarbeiten der zugrundeliegenden analytischen Modellierung eines physikalischen Problems – um eine wissenschaftliche bzw. experimentelle Arbeitsweise handelt, denn computersimulierte Experimente unterscheiden sich von Realexperimenten in einer Reihe von Charakteristika, wie die folgende Auflistung zeigt.

DATENGENERIERUNG VS. DATENREPRODUKTION Ein Messwert ist ein von einer oder mehreren Variablen abhängiger Wert einer Messgröße aus einem Kontinuum von weiteren Messwerten, der durch die Durchführung einer Messung generiert wird. Im Gegensatz dazu werden in Simulationen „Messwerte“ aufgrund eines bereits in das Programm implementierten Modells eines physikalischen Phänomens reproduziert, indem Ausgangsbedingungen in eine Modellgleichung eingesetzt werden. Die Lösung dieser Modellgleichung produziert dann einen „Messwert“. In vielen Simulationsumgebungen werden Daten aber als absolut und „gültig“ präsentiert: So werden oftmals unmittelbar numerische Werte ausgegeben. Das Ablesen eines Messinstruments entfällt und es wird somit ein „Messwert“ beliebig hoher Präzesion suggeriert, die es in realen Experimenten nicht gibt.

UNSICHERHEIT DER MESSUNG Der Umgang mit Messunsicherheiten bei der experimentellen Erarbeitung von Naturphänomenen ist elementarer Bestandteil der Physik (Heinicke, 2012; Hellwig, 2012). In der Natur ist ein einmal gemessener Messwert nicht in beliebiger Präzision zu reproduzieren. Eine Reihe von Messwerten, die unter gleichen Bedingungen aufgenommen wurden, weisen daher i. d. R. eine Streuung auf. Die Unsicherheit ist dann „ein mittleres Maß der Streuung einer Datenmenge“ (Heinicke, 2012, S. 93). Im Gegensatz dazu weisen Messwerte aus Simulationen i. d. R. keine Unsicherheit auf.

MENGE DER MÖGLICHKEITEN Oft unterscheiden sich Computersimulationen von realen Experimenten in der Menge der zur Verfügung stehenden experimentellen Möglichkeiten. Experimentelle Handlungsmöglichkeiten sind im Hinblick auf die Kontrolle von Variablen, Möglichkeiten der Veränderungen am Aufbau sowie beim Umgang mit Fehlern eingeschränkt. Letzteres ist

insbesondere im Hinblick auf die Analyse *möglicher* Fehlerquellen von Bedeutung, denn zur Evaluation experimenteller Daten und Beobachtungen gehört auch der Ausschluss möglicher einflussnehmender Faktoren.

DER FAKTOR ZEIT Reale physikalische Experimente benötigen Zeit. Die Planung und der Aufbau eines Experiments entfallen bei einer Simulation meist völlig. Auch der eigentliche Messprozess wird extrem verkürzt dargestellt: Wenige Mausklicks reichen, um weitere Messreihen aufzunehmen. Dies ist bei Realexperimenten grundsätzlich anders.

DIMENSIONALITÄT Die Schnittstelle zwischen Lernenden und computergestützten Simulationen bildet ein Bildschirm, der das Geschehen zweidimensional darstellt. Im Gegensatz dazu sind Realexperimente per se dreidimensional zu erfassen. Reale Experimente können daher taktile Erfahrungen bieten, die nach der Theorie der Embodied Cognition die Aneignung von konzeptuellen Wissen fördern können (de Jong, Linn & Zacharia, 2013).

VEREINFACHUNG VS. KOMPLEXITÄT Simulationen bilden die Realität elementarisiert ab, da in den Computerprogrammen i. d. R. Elementarisierungen und Näherungen verwendet werden. Im Gegensatz dazu liegt bei Realexperimenten eine höhere Komplexität vor. In der vorliegenden Arbeit wird ein Experiment zum Fadenpendel verwendet. Bei der physikalischen Modellierung dieses Phänomens müssen eine Reihe von Näherungen getroffen werden, z. B. die Kleinwinkelnäherung, Näherungen zur Ausdehnung der angehangenen Masse oder zum Einflusses der Auslenkung auf die Periodendauer (für eine genauere Darstellung zu den Annahmen bei der Modellierung des Fadenpendels siehe Abschnitt 4.1.2).

UNTERSCHIEDE IN DER HANDHABUNG Simulationsexperimente bedürfen eines geeigneten Eingabegerätes. Das Nutzen von Computermaus bzw. Tastatur zur Manipulation eines experimentellen Aufbaus ist grundsätzlich verschieden von der Manipulation realer Objekte. Realexperimente zeichnen sich durch eine direkt erfahrbare Bedienbarkeit und haptische Wahrnehmung aus.

MOTIVATIONALE ASPEKTE Nach der Selbstbestimmungstheorie der Motivation (Deci & Ryan, 1993) lassen sich u. a. Kompetenzerleben und Autonomie als psychische Grundbedürfnisse und „treibende Kräfte“ im Streben nach persönlicher Entwicklung ableiten. Es kann angenommen werden, dass diese beiden Faktoren in den verschiedenen Lernsettings unterschiedlich ausgeprägt sind und daher Einfluss auf motivationale Aspekte beim

> Lernen nehmen. In einer Simulationsumgebung ist die Autonomie insofern eingegrenzt, als dass lediglich der Handlungsspielraum zugelassen ist, den ein Programmierer im Zuge der Gestaltung des Programms vorgesehen hat. Im Gegensatz dazu ist der Lernende in einem realen Experiment befähigt, auch eigene, neue oder unkonventionelle Wege zu bestreiten. Empirische Ergebnisse untermauern diese These. Corter, Esche, Chassapis, Ma und Nickerson (2011) konnten zeigen, dass der Umgang mit realen Experimenten zu einer höheren Motivation und einer längeren Beschäftigung mit der Aufgabe führen kann. Es ist aber auch denkbar, dass durch den eingeschränkten Handlungsspielraum die Komplexität reduziert wird (s. o.), so dass z. B. durch einen *scaffolding*-Effekt das Autonomieerleben bzw. die Motivation steigt. Auch für diese Argumentation gibt es empirische Hinweise (z. B. Kebritchi, Hirumi & Bai, 2010).

Aufgrund der angeführten Unterschiede zwischen Computer- und Realexperimenten ist es fraglich, ob es möglich ist, Ergebnisse, die in einer virtuellen „Experimentierumgebung beobachtet werden, didaktisch sinnvoll auf realen naturwissenschaftlichen Unterricht, insbesondere Physikunterricht übertragen zu können“ (Künsting et al., 2008, S. 2). Bisherige Arbeiten zur Einflussnahme des Mediums auf Experimentierprozesse weisen divergierende Befunde auf. Es gibt zum einen Untersuchungen, die keine Unterschiede zwischen Real- und Simulationsexperimenten zeigen. Im Hinblick auf Lernleistungen wurde dies von Brell (2008), Brell, Theyßen, Schecker und Schumacher (2006), Klahr, Triona und Williams (2007), Neugebauer (2006), Renken und Nunez (2013), Sander, Schecker und Niedderer (2003), Zacharia und Constantinou (2008) sowie Zacharia und Olympiou (2011) berichtet. Im Hinblick auf das Problemlösen wurde dies von Neugebauer (2006) berichtet. Auch die Anwendung der Variablenkontrollstrategie bei der Planung von Experimenten wird nicht durch das Medium beeinflusst (Triona & Klahr, 2003).

Es gibt aber auch empirische Arbeiten, die Unterschiede zwischen realen und virtuellen Umgebungen beim Lernen von Naturwissenschaften finden. So finden sich positive Effekte für das computergestützte Experimentieren z. B. hinsichtlich der Effizienz (Brell, 2008; Finkelstein et al., 2005) sowie hinsichtlich des konzeptuellen Verständnisses (Zacharia, 2007; Zacharia, Olympiou & Papaevripidou, 2008). Zacharia, Loizou und Papaevripidou (2012) berichten in einer mit Lernenden im Vorschulalter durchgeführten Studie zum Thema Balkenwaage, dass die Körperlichkeit (engl. *physicality*) von realen Experimenten den Erwerb konzeptuellen Wissens eher fördert als eine virtuelle Lernumgebung.

Darüber hinaus gibt es mehrere Reviews auf Metaebene zu diesem Thema. Smetana und Bell (2012) fassen die Ergebnisse von 61 empirischen Vergleichsstudien bzgl. Computer- und Realexperimenten

zusammen. Der überwiegende Teil der in dieser Metastudie eingeschlossenen Arbeiten berichtet von positiven Effekten beim Einsatz von Computerexperimenten beim Erwerb von Fachwissen und prozeduralem Wissen (49 von 61 Studien). Bei elf der Studien konnte kein Unterschied zwischen realem und virtuellem Lernsetting gefunden werden. In einem weiteren Review sind Rutten, van Joolingen und van der Veen (2012) der Frage nachgegangen, inwiefern traditionelle Lernszenarien durch Computersimulationen erweitert und verbessert werden können. Die 51 zur Analyse herangezogenen Arbeiten berichten überwiegend positive Effekte durch den Einsatz von Computersimulationen. Dabei ist jedoch zu beachten, dass in dieses Review nicht nur klassische Vergleichsarbeiten (real vs. virtuell) eingingen, sondern überwiegend Studien Berücksichtigung fanden, die Interventionen zu Instruktionsstrategien, Klassenraum-Szenarien sowie verschiedenen Arten der Visualisierung beim Lernen mit Computersimulationen untersuchten.

Die beiden Reviews bieten eine Möglichkeit, die Art der untersuchten abhängigen Variablen zu analysieren. Dazu werden im Folgenden nur die in die Metaanalysen eingegangenen Studien mit Physikbezug betrachtet. In das Review von Smetana und Bell (2012) sind 16 Studien eingegangen, die einen physikalischen Inhalt als Lerngegenstand hatten. Davon haben die Autoren bei sechs dieser Studien abhängige Variablen aus dem Bereich Fachwissen („content knowledge“) identifiziert. Je vier Studien untersuchten abhängige Variablen aus dem Bereich Konzeptwechsel und aus dem allgemeinpädagogischen Bereich (die Autoren erläutern die vorgenommene Unterscheidung bzgl. Fachwissen und Konzeptwechsel leider nicht). Lediglich zwei Studien legen ihren Fokus auf prozedurale Variablen („process skills“) beim Experimentieren mit Computersimulationen (vgl. Smetana und Bell, 2012, S. 1343, S. 1348, S. 1351, S. 1353). In das Review von Rutten et al. (2012) sind acht Studien mit Physikbezug eingegangen. Eine Studie untersucht motivationale Aspekte, die übrigen sieben ebenfalls den Erwerb konzeptuellen Verständnisses (vgl. Rutten et al., 2012, S. 139, S. 142, S. 145f, S. 149). Die Analyse der beiden Reviews zeigt sehr deutlich, dass viele Studien oft abhängige Variablen in die Untersuchung einschließen, eher auf den Erwerb konzeptuellen Wissens fokussieren.

Zusammengefasst finden sich in der Literatur keine eindeutigen Befunde zum Einfluss des Mediums auf Lernprozesse beim Experimentieren. Dies ist auch der Tatsache geschuldet, dass zum einen unterschiedliche abhängige Variablen in Zusammenhang mit dem Experimentieren in die Untersuchungen eingeschlossen wurden und diese zum anderen unterschiedlich operationalisiert wurden. Zudem zeigt sich eine starke Abhängigkeit von Kontext und der eigentlichen Aufgabenstellung (Rutten et al., 2012).

Bezogen auf die Fragestellung der vorliegenden Arbeit gibt es gute Gründe, einen Einfluss der Lernumgebungen allein aufgrund ihrer Unterschiedlichkeit auf Ebene der Experimentierprozesse anzunehmen, der für die Argumentation der Probanden von Bedeutung sein kann.

2.3 EIGENE VORARBEITEN ZUM ARGUMENTIEREN BEIM EXPERIMENTIEREN: WELCHE ARGUMENTE ENTWICKELN LERNENDE ANHAND EXPERIMENTELLER DATEN UND BEOBACHTUNGEN?

2.3.1 *Forschungsfrage*

In Abschnitt 2.1.4 wurde dargelegt, dass es lediglich eine überschaubare Anzahl an naturwissenschaftsdidaktischen Studien gibt, die konkret das Argumentieren im Kontext des naturwissenschaftlichen Experimentierens untersucht haben. Bisher ungeklärt ist insbesondere die Frage, welche Argumente Schüler auf der Basis von selbstständig erhobenen Messdaten und experimentellen Beobachtungen entwickeln. Diese Fragestellung wurde anhand des Experiments zum Fadenpendel in Vorarbeiten des Autors untersucht (vgl. Ludwig, 2011). Im Rahmen der vorliegenden Arbeit wurden die Daten erneut analysiert und die Studie in einem zweiten physikalischen Kontext durchgeführt (Lau, 2013). Da die Ergebnisse aus diesen Vorarbeiten von zentraler Bedeutung für die vorliegende Arbeit sind, sollen nach einem kurzen Überblick über die Methode die wesentlichen Ergebnisse aus diesen Vorarbeiten dargestellt und diskutiert werden.

Im Rahmen der Vorarbeiten stand im Zentrum des Interesses die folgende Forschungsfrage:

> Welche Argumente verwenden Lernende für das Beibehalten bzw. Verwerfen von selbst aufgestellten Hypothesen bei Evaluation selbst generierter experimenteller Daten und Beobachtungen, und wie lassen sich diese Argumente kategorisieren?

2.3.2 *Untersuchungsdesign*

Die angeführte Forschungsfrage wurde in einem Studiendesign untersucht, das weitestgehend mit dem Design der Hauptuntersuchung der vorliegenden Arbeit vergleichbar ist (vgl. Abschnitt 4.4): Nach dem Aufstellen einer physikalischen Hypothese (die mit einer hohen Wahrscheinlichkeit fachlich inkorrekt ist) wurden die Probanden aufgefordert, diese Hypothese experimentell zu überprüfen. Im Anschluss an die Experimentierphase wurde zunächst erfasst, ob die vor dem Experiment aufgestellte Hypothese verworfen oder beibehalten

wurde. Falls die Hypothese verworfen wurde, wurde zudem gefragt, welche neue Hypothese auf der Grundlage der Experimentierergebnisse aufgestellt wurde. In einem offenen Interview wurden dann die Argumente für das Beibehalten oder Verwerfen der eingangs aufgestellten Hypothese erarbeitet, indem die folgende Frage gestellt wurde: „Warum verwirfst / behältst du deine Vermutung von Beginn bei?“ Dieser Stimulus zur argumentativen Auseinandersetzung mit den Experimentierergebnissen ist nötig, da dieser Prozess bekanntermaßen von Lernenden nur selten allein aufgegriffen wird (vgl. Abschnitt 2.1.4). Die Interviewer haben dann ggf. mit weiteren Stimuli und Nachfragen versucht, die Darstellung des Probanden weiter anzuregen (z. B. mit Fragen wie „Wie überzeugt bist du, dass es richtig ist, dass du deine Vermutung verwirfst / beibehältst?“).

Aus Gründen der Lesbarkeit sei daher an dieser Stelle für eine detaillierte Darstellung der verwendeten Methodik auf die entsprechenden Abschnitte verwiesen. Dazu gehören insbesondere die Argumentation hinsichtlich der Auswahl der physikalischen Themen, an denen experimentelle Daten und Beobachtungen erzeugt werden können, die mit einem fachlich inkorrekten Präkonzept verbunden sind (Abschnitt 4.1), die Beschreibung der computerbasierten Simulation sowie des Realexperiments zum Fadenpendel (Abschnitt 4.1.4) sowie der Ablauf der Untersuchung (Abschnitt 4.5).

Neben einer computersimulierten sowie einer realen Experimentierumgebung zum Fadenpendel wurde – im Gegensatz zur Hauptuntersuchung – zusätzlich ein Beispiel aus der Wärmelehre verwendet, bei dem die Temperatur in Festkörpern unterschiedlicher Größe bestimmt werden sollte. Dies geschah zum einen mit dem Ziel, eine möglichst hohe Vielfalt an Argumenten zu generieren („sampling for variation“, Morse & Niehaus, 2009, S. 62), zum anderen sollte so die Abhängigkeit der Argumente von spezifischen Charakteristika der Inhalte bzw. Kontexte geprüft werden. Das Beispiel aus der Wärmelehre basiert auf einem weit verbreiteten Präkonzept, dem zufolge gewöhnliche Materialien in der Lage sind, „aktiv“ zu wärmen; ein in einen Wollpullover eingewickeltes Thermometer soll also eine höhere Temperatur anzeigen als eines, das nicht umhüllt ist (Albert, 1978; Duit, 1986). Dieses Präkonzept wurde zunächst empirisch überprüft („Wie beeinflusst die Umhüllung eines Thermometers mit einem Wollpullover die angezeigte Temperatur?“). Dabei stellten rund 70 % der Probanden eine falsche Hypothese auf (Decker, 2012). Auch bei diesem Beispiel werden Probanden beim Experimentieren daher mit nicht-hypothesenkonformen Daten konfrontiert (vgl. Abschnitt 4.1). Für das Experiment wurden einfache, schulübliche Materialien zur Verfügung gestellt (siehe Abbildung 8 in Anhang A).

Insgesamt führten die Probanden so drei verschiedene Experimente durch (Realexperiment zum Fadenpendel, Computersimulation zum

Fadenpendel sowie Temperaturmessung in Festkörpern mit realen Materialien).

2.3.3 *Stichprobe*

An der Studie nahmen 129 Schüler aus vier verschiedenen Gymnasien in Nordrhein-Westfalen und Berlin teil. Das Alter lag zwischen 12 und 18 Jahren, im Mittel bei 14.6 Jahren ($SD = 1.1$ Jahre). Die Probanden wurden randomisiert den drei Gruppen zugeordnet (38 Probanden arbeiteten mit dem Realexperiment zum Fadenpendel, 41 mit der Computersimuation zum Fadenpendel sowie 48 mit dem Temperaturexperiment). Die Interviews dauerten im Mittel $\bar{t} = 3\,\mathrm{min}\,57\,\mathrm{s}$ ($SD = 45\,\mathrm{s}$). Die Interviews wurden zur Auswertung transkribiert.

2.3.4 *Analyse der Interviews*

Anhand der transkribierten Interviews konnten in Hinblick auf die Forschungsfrage die von den Probanden gegebenen Argumentationen für das Verwerfen bzw. Beibehalten von selbstaufgestellten Hypothesen bei unerwarteter Datenlage analysiert werden. Alle Kategorien sind so konstruiert, dass sowohl Äußerungen für als auch gegen die eingangs aufgestellte Hypothese zugeordnet werden können. Die Kategorie Expertenwissen beispielsweise zeigt dies sehr eindrucksvoll. Die Äußerung *Es müsste mir noch jemand sagen, dass das stimmt* kann sowohl als Argument für als auch gegen eine Eingangshypothese auftreten. Zur Erstellung eines Kategoriensystems wurden zunächst aus Einzelfällen induktiv zehn Kategorien abgeleitet (Krüger & Riemeier, 2014). Im Folgenden soll dieser Kodierprozess für einige dieser Kategorien exemplarisch mit Datenmaterial belegt werden. Dazu sind jeweils Interviewexzerpte dargestellt, im Anschluss wird erläutert, wie die entsprechenden Textstellen zur Kategorienbildung geführt haben:

Schülerin, 14 Jahre alt, 8. Klasse, Fadenpendel, Realexperiment, falsche Eingangshypothese. (Auf Füllworter und Halbsätze wurde zur besseren Lesbarkeit verzichtet)

Interviewer:	Du hast jetzt gerade experimentiert. Behältst du deine Vermutung von Beginn an bei?	1 2
Schülerin:	Hmm, ja.	3
Interviewer:	Warum behältst du deine Hypothese bei?	4
Schülerin:	Da kam halt raus, desto schwerer das Pendel ist, desto länger dauert es. Das war nicht viel, nur so hundertstel Sekunden. Das eine Mal kam irgendwie 2.97 s bei 2 kg	5 6 7

und das andere mal, als ich viel weniger genommen hatte, kam nur 2.96 raus. Nur so minimal, aber. [...]

Aus dieser Äußerung wurden zwei Argumente abgeleitet. Zum einen bezieht sich die Schülerin auf die experimentell erhobenen Messdaten und zieht diese als Evidenz heran (Zeile 7–9). Eine solche Äußerung wurde in diesem und in weiteren Interviews der Argumentkategorie *Daten als Evidenz* zugeordnet. Die Schülerin erläutert weiterhin, dass die gefundenen Unterschiede sehr klein waren (Zeile 6 und 9). Diese Äußerung steht exemplarisch für Äußerungen, die der Argumentkategorie *Messunsicherheiten (implizit)* zugeordnet wurden.

Schülerin, 13 Jahre alt, Computersimulation zum Fadenpendel, falsche Eingangshypothese

Interviewer: Du hast gerade das Experiment durchgeführt. Behältst du deine Vermutung von Beginn bei?

Schülerin: Nein.

Interviewer: Warum verwirfst du deine Vermutung?

Schülerin: [...] Ich habe angenommen, dass, je größer die Masse, desto länger die Zeit, die das Pendel schwingt. Weil der Luftwiderstand dann größer ist. Aber das hat damit nichts zu tun, oder? In meinen Ergebnissen war es dasselbe [die Schwingungsdauer], egal wieviele Gewichte ich benutzt habe. Ich habe einmal mit 0.1 kg gemessen und einmal mit 2 kg und es war immer dasselbe. Immer 2.958 oder sowas.

Interviewer: Wie sicher bist du dir mit deiner Entscheidung?

Schülerin: Ich bin todsicher.

Interviewer: Warum?

Schülerin: Weil ich etwas völlig anderes beobachtet habe, als ich das Experiment durchgeführt habe. Der Computer wird seine Gründe dafür haben.

Interviewer: Kannst du erklären warum?

Schülerin: Ja, die Simulation wurde von einer Universität programmiert. Das muss dann richtig sein.

Aussagen aus diesem Interview wurden in drei Argumentkategorien eingeordnet. Zum einen zieht die Schülerin *Daten als Evidenz* heran (Zeile 12–14). In Zeile 18–22 verweist die Schülerin darauf, dass die Simulation als solche bereits Expertise enthält. Dabei beruft sie sich darauf, dass die Simulation von einer Universität entwickelt wurde und

daher „richtige“ Daten liefert (Anmerkung: Die Simulation enthielt ein Logo der Universität. Zudem waren die Probanden darüber aufgeklärt, dass es sich um eine wissenschaftliche Studie handelt). Diese und ähnliche Aussagen in weiteren Interviews wurden daher der Kategorie *Expertenwissen* zugeordnet. Diese Kategorie zeichnet sich dadurch aus, dass hier auf eine bereits in das Experiment implementierte Expertise verwiesen wird. Im letzten Satz wird zudem deutlich, dass die Schülerin in ihrer Begründung *ad hoc* ein physikalisches Konzept heranzieht, um den Hypothesenwechsel zu begründen. Da diese ad hoc heran gezogenen fachlichen Konzepte nicht zwingend richtig sein müssen, werden Aussagen dieser Art in die Argumentkategorie *Heranziehen einer (falschen) physikalischen Theorie* eingeordnet.

2.3.5 *Ergebnisse*

Im Folgenden sind für jede Argumentkategorie, die durch das oben dargelegte Verfahren abgeleitet werden konnte, eine Beschreibung und ein Beispiel aufgeführt. Jede Kategorie repräsentiert dabei einen möglichen Argumenttyp, der in Argumentationen sowohl für als auch gegen einen Hypothesenwechsel auftreten kann.

INTUITION Diese Kategorie enthält Aussagen, die auf eine intuitive, gefühlsgeleitete Auseinandersetzung mit der Experimentiersituation zurückzuführen sind. Beispiel für eine Schüleraussage, die dieser Argumentkategorie zugeordnet wird: „Das Gefühl sagt halt, dass es, wenn die Masse größer wird, heftiger schwingt. . . . “

IGNORANZ Diese Kategorie liegt bei Nennung des Experimentierergebnisses bei gleichzeitiger Argumentation für eine in Bezug auf die Datenlage widersprüchliche Hypothese vor. Beispiel: Der Proband trifft folgende Aussage: „Nach dem Experiment hab ich keine Unterschiede gemerkt, als ich die Gewichte dazu getan habe.“ Dennoch gibt er als neue Hypothese nach dem Experimentieren an, das die Schwingungsdauer mit zunehmender Masse kleiner wird.

EXPERIMENTIERKOMPETENZ (ZENTRAL) Hierzu zählen Aussagen, die sich in Bezug auf die eigene Person aus einer technischen bzw. handwerklichen Perspektive mit dem zuvor durchgeführten Experimentierprozess auseinandersetzen. Beispiel: „Kann ja sein, dass ich irgendwie falsch gestoppt habe, also bei dem einen zu früh, bei dem anderen zu spät.“

EXPERIMENTIERKOMPETENZ (PERIPHER) Aussagen, die auf dem physikspezifischen Selbstkonzept bzw. der Selbstwirksamkeitserwartung hinsichtlich des Experimentierens basieren, werden dieser Kategorie zugeordnet. Beispiel: „Ja, ich weiß nicht, das

ist immer noch so der Hintergedanke, dass das ja noch falsch sein könnte, dass ich da noch irgendwas falsch gemacht habe."

EIGNUNG DES EXPERIMENTS Aussagen, welche die fehlerfreie Funktion des experimentellen Aufbaus bzw. die fehlerfreie Programmierung oder generell die Eignung des Aufbaus bzw. der Simulation z. B. hinsichtlich Variablenkontrollmöglichkeiten oder idealisierter Randbedingungen infrage stellen, werden in dieser Kategorie erhoben. Beispiel: „Ich weiß nicht, weil ich mir eigentlich doch ziemlich sicher war mit meiner Vermutung, aber dann doch etwas völlig anderes raus kam, was für mich nicht so ganz schlüssig ist. ... Der Computer hatte ja immer das gleiche Ergebnis, aber ich hatte halt gedacht, dass wenn man die Masse verändert, dass da auch unterschiedliche Werte rauskommen. Es war dann ja nicht so der Fall. Eigentlich müsste ich ja dem Computer ... Ich bin mir jetzt noch ein bisschen unsicher, weil solche anderen Faktoren ... nicht berücksichtigt wurden."

HERANZIEHEN EINER (FALSCHEN) PHYSIKALISCHEN THEORIE Diese Kategorie umfasst Argumente, die das Experimentierergebnis durch Heranziehen einer (oftmals falschen) physikalischen Theorie erklären wollen. Eine Auseinandersetzung mit den nicht-hypothesenkonformen Daten wird somit vermieden. Es handelt sich dabei i. d. R. um nicht-wissenschaftliche Ad-hoc-Erklärungen. Beispiel: „Je mehr Gewicht da dran ist, desto kleiner die Zeit, denn das Gewicht verstrammt den Faden. Und dadurch kommt der schneller zurück."

EXPERTENWISSEN In dieser Kategorie werden Argumente zusammengefasst, die sich explizit auf das vorhandene Expertenwissen stützen oder nach einer externen Bestätigung verlangen. Ebenso umfasst diese Kategorie Antworten, die zwar den Hypothesenwechsel vertreten, aber von den Probanden aufgrund mangelnden Zuspruchs seitens eines Experten (z. B. Lehrers) skeptisch beurteilt werden. Beispiel eines Probanden, der mit der Computersimulation gearbeitet hat: „Ich hab das halt ausprobiert und das blieb gleich. ... Die Experimente sind ja da alle schon geprüft worden. Dann geh ich davon aus, dass das alles richtig ist. Die benutzen bestimmt auch Formeln mit denen man das alles ausrechnen kann." Weiteres Beispiel: „Es müsste mir noch jemand sagen, dass das stimmt. Das liegt jetzt also nicht am Experiment."

MESSUNSICHERHEITEN (IMPLIZIT) Dieser Kategorie werden Aussagen zugeordnet, die sich durch eher intuitive Interpretationen der Daten auszeichnen. Ursächlich für die Verwendung dieser Kategorie ist vermutlich nicht vorhandenes Wissen über Messunsicherheiten. Kennzeichnendes Merkmal ist dabei, dass auf

Unsicherheiten in den Messdaten nur implizit Bezug genommen wird. Beispiel: „Dann hab ich es auf 40° gestellt und das Gewicht erhöht und mit der Stoppuhr getestet und dann war es einmal 2,95 und einmal 2,99. Das Pendel braucht also länger."

MESSUNSICHERHEITEN (EXPLIZIT) Diese Argumentkategorie umfasst Aussagen, welche die Relevanz der Betrachtung von Messunsicherheiten in den Daten explizit benennen. Diese Aussagen zeugen davon, dass der Proband in seiner Entscheidung Messunsicherheiten berücksichtigt hat. Beispiel: „Es kann ja natürlich Messungenauigkeiten geben. ... Es kann natürlich auch sein, dass das jetzt Zufall war, dass das [die Schwingungsdauer] immer gleich war."

DATEN ALS EVIDENZ Diese Argumentkategorie umfasst alle Aussagen, die Messdaten oder experimentelle Beobachtungen als Evidenz heranziehen. Die Messdaten und Beobachtungen können dabei implizit oder explizit berücksichtigt werden. Ein Beispiel für eine explizite Berücksichtigung findet sich in den Interviewexzerpten oben. Eine Schüleräußerung für ein implizites Nutzen von Beobachtungen und Messdaten als Evidenz: „Es wurde ja durch das Experiment bewiesen, dass die Zeit immer gleich war."

Mit den vorliegenden Experimentiersituationen konnten insgesamt zehn verschiedene Argumentkategorien gefunden werden, die beim Umgang mit nicht-hypothesenkonformen Daten verwendet wurden.

Die Argumentkategorien wurden aus Validitätsgründen nicht im Hinblick auf die Häufigkeit einer Nennung analysiert, da die geführten Interviews mit dem Ziel durchgeführt wurden, möglichst viele verschiedene Argumente zu sammeln und nicht auszuschließen ist, dass es zu Verzerrungen kommt (eine genauere Darstellung dieses Problems findet sich in Abschnitt 4.3.1). Für die Daten der Stichprobe, die mit dem Fadenpendel gearbeitet hat, wurde jedoch ausgezählt, wie viele Äußerungen insgesamt einer Argumentkategorie zugeordnet wurden. In den Interviews mit den 81 Probanden fanden sich 157 Äußerungen, die einer der zehn Kategorien zugeordnet wurden.

Um die eindeutige Zuordnung der Antworten in die Kategorien sicher zu stellen, wurden die Interviews durch zwei unabhängige Rater kategorisiert. Anhand eines Kodiermanuals mit den Kategorienbeschreibungen und einer Schulung bewerteten die Experten die Aussagen aller Probanden auf einer dreistufigen Ratingskala (trifft nicht zu, trifft teilweise zu, trifft voll zu) zum Auftreten einer Argumentkategorie. Diese Zuordnungsstufen weisen Ordinalskalenniveau auf. Als Maßzahlen für die Reliabilität ordinalskalierter Ratings sind u. a. Rangkorrelationskoeffizienten geeignet (Gwet, 2010; Wirtz & Caspar, 2002). Zur Bestimmung der Reliabilität der Raterurteile wurde daher

die Stärke des Zusammenhangs der Beurteilungen durch den Korrelationskoeffizienten Spearmans ρ bestimmt. Weiterhin wurde die prozentuale Übereinstimmung bestimmt. Die über alle zehn Kategorien gemittelte prozentuale Übereinstimmung ergibt sich zu 91.4 % (Fadenpendel) bzw. 85.4 % (Temperatur im Festkörper); die gemittelte Korrelation zu $\rho = .80$ (Fadenpendel) bzw. $\rho = .61$ (Temperatur im Festkörper). Eine detaillierte Aufstellung der Beurteilerreliabilität bzw. -übereinstimmung je Kategorie ist in Tabelle 22 in Anhang A dargestellt. Die Reliabilitätsanalyse zeigt daher, dass die Einordnung von Schüleraussagen relativ unabhängig vom Beurteiler ist und zudem mit einer hinreichenden Reliabilität erfolgt. Es ist daher davon auszugehen, dass die Beschreibung der Kategorien eine genügende Präzision und Trennschärfe aufweist.

2.3.6 *Diskussion ausgewählter Ergebnisse*

Im weiteren Verlauf der Arbeit sind vier der zehn berichteten Argumentkategorien zur Untersuchung der Hauptforschungsfragen herangezogen worden (für eine Begründung zu dieser Entscheidung siehe Abschnitt 2.5.3). Es sollen daher die Argumentkategorien Intuition, Expertenwissen, Messunsicherheiten (explizit) sowie Daten als Evidenz ausführlicher diskutiert werden. Darüber hinaus wird die Kategorie Messunsicherheiten (implizit) diskutiert, was zu einer stärkeren Differenzierung zwischen den beiden Kategorien, die implizit bzw. explizit Messunsicherheiten in Daten berücksichtigen, beitragen soll. Auch wenn der nachfolgende Abschnitt Literatur aus Wissenschafts- und Argumentationstheorie heranzieht, werden die Argumentkategorien allenfalls überblicksartig und mit einer Fokussierung auf die Physik bzw. auf den Physikunterricht diskutiert. Es sei hier auf den umfangreichen Literaturkorpus aus der Argumentationstheorie verwiesen (z. B. Walton, 1995, 2008, 2010, 2016).

INTUITION Das Auftreten von intuitiven Prozessen in den Naturwissenschaften ist weithin bekannt (Bunge, 1962). Insbesondere dann, wenn es um statistische Entscheidungen geht, die mit einer Unsicherheit behaftet sind, neigen Menschen zu intuitiven Entscheidungen, wie Kahneman und Tversky (1973, S. 273) beschreiben:

> In making predictions and judgements under uncertainty people do not appear to follow the calculus of chance or the statistical theory of prediction. Instead, they rely on a limited number of heuristics, which sometimes yield reasonable judgments and sometimes lead to severe and systematic errors.

Auch im vorliegenden Beispiel zum Fadenpendel ist es notwendig, eine statistische Entscheidung zu treffen. Auf der Grundlage des

(idealerweise korrekt) durchgeführten Experiments gilt es, bei Variation der angehangenen Masse die Gleichheit der Mittelwerte der Schwingungsdauer bei verschiedener Pendelmasse zu testen (Anmerkung: Es wurde von den Probanden nicht erwartet, die Gleichheit der Mittelwerte inferenzstatistisch zu prüfen. Es gibt geeignete alternative Verfahren für die Sekundarstufe, die auf das Abschätzen der maximalen Abweichung vom Mittelwert abzielen z. B. Hellwig, 2012). Die Bezugnahme auf intuitive Prozesse in den Interviews ist daher nicht verwunderlich. Deshalb wird gefordert, intuitive Prozesse beim Lernen allgemein (Fischbein, 1987) bzw. beim Lernen von Naturwissenschaften zu berücksichtigen (Fensham & Marton, 1992). In den Naturwissenschaftsdidaktiken geschieht dies bisher jedoch nur in wenigen Arbeiten (z. B. im Bereich des Bewertens von Dittmer & Gebhard, 2012). Generell erweist es sich als durchaus schwierig, Menschen von intuitiven Prozessen abzubringen (Sedlmeier, Lovett & Priti, 2007).

In der vorliegenden Arbeit wurden Aussagen der Kategorie Intuition zugeordnet, die sich durch eine „gefühlsgeleitete" Auseinandersetzung mit den experimentellen Daten und Beobachtungen auszeichnen. Dieses Vorgehen ist insbesondere vor dem Hintergrund von Definitionen zur Intuition aus der Kognitionspsychologie bzw. den Verhaltenswissenschaften gerechtfertigt: „[T]he notion of intuition is invariably related to feelings and emotions" (Hogarth, 2001, S. 60). In den gängigen Definitionen der Intuition lassen sich aber zudem auch Merkmale von Intuition finden, die im Datenmaterial dieser Studie nicht gefunden werden konnten. So entsteht Intuition zudem oftmals instantan in einer Situation (Hodgkinson, Langan-Fox & Sadler-Smith, 2008) bzw. schnell (Gigerenzer & Gaissmaier, 2011). Intuitive Prozesse laufen zudem zwar unbewusst ab (Dane & Pratt, 2007; Hodgkinson et al., 2008), die Ergebnisse dieser Prozesse werden jedoch bewusst wahrgenommen und können sich als „Ahnung" (engl. *hunch*) bzw. „Bauchgefühl" äußern (Dane & Pratt, 2007; Hodgkinson et al., 2008). An dieser Stelle sei angemerkt, dass der Literaturkorpus das Konstrukt Intuition mit einem deutlich höheren Auflösungsgrad beschreibt, z. B. hinsichtlich der Vergleichbarkeit von Heuristiken und Intuiton (Keller, Bohner & Erb, 2000) und zudem verwandte Konstrukte deutlicher abgrenzt (siehe z. B. Hodgkinson et al., 2008). Die Verwendung des Begriffs Intuition in der vorliegenden Arbeit beruht daher nur auf einem bestimmten Teilaspekt gängiger Definitionen.

EXPERTENWISSEN Die Argumentkategorie Expertenwissen weist Parallelen zu dem aus der Logik bekannten *argumentum ad verecundiam* auf, welches bereits 1690 von John Locke erstmals beschrieben wurde (Walton, 2010). Darunter wird das Vorgehen verstanden, in einer Argumentation auf eine Person zu verweisen, die in ihrem Feld eine gewisse Expertise bzw. kognitive Autorität aufzuweisen hat, sodass es „unsittlich" wäre („breach of modesty", Walton, 2010, S. 47),

dieses anzuzweifeln, weil damit die allgemein anerkannte Expertise dieser Person negiert werden würde. Dieser Typus des Arguments kann in besonderen Fällen als unangemessen gelten: Zum einen können Trugschlüsse entstehen, indem trotz gegenläufiger Evidenz durch einen Verweis auf Expertise für eine andere Hypothese argumentiert wird (Hansen, 2015), zum anderen ist es oftmals schwierig festzustellen, wie die Expertise anderer beurteilt werden kann (Walton, 2016). Entsprechend konstatiert Walton (2016) treffend: „Argument from expert opinion has always been a form of reasoning that is on a razor's edge" (S. 118). Im Rahmen von Schule – eine institutionalisierte Lehr-Lern-Situation– haben Schüler gute Gründe, sich auf Expertise (z. B. Lehrkräfte) zu verlassen und darauf in epistemischen Prozessen wie dem Argumentieren zu verweisen. Es ist für Lernende unmöglich, diese Expertise stets zu überprüfen oder infrage zu stellen. Im Kontext des Experimentierens bedeutet dies beispielsweise, dass Schüler sich auf die Richtigkeit einer analytischen Lösung eines physikalischen Problems in einer verwendeten Computersimulation oder die Eignung eines (vorgelegten) experimentellen Aufbaus verlassen können müssen.

MESSUNSICHERHEITEN (IMPLIZIT) Es wurden zwei Kategorien identifiziert, die – explizit oder implizit – einen Bezug zu den Unsicherheiten der Messdaten aufweisen. Die Argumentkategorie Messunsicherheiten (implizit) ist charakterisiert dadurch, dass Schüler auf der Grundlage von nicht vorhandenem Wissen über Messunsicherheiten für eine Hypothese argumentieren, die der Datenlage – bei fachlich korrekter Analyse der Messunsicherheit – widerspricht, indem z. B. nur eine unbedeutende Variation in den Daten interpretiert wird. Hierbei handelt es sich aus Sicht der Schüler um eine formallogisch korrekte Argumentation. Wird dieses Vorgehen jedoch aus fachphysikalischer Perspektive betrachtet, handelt es sich hierbei um einen logischen Fehlschluss, der vergleichbar ist mit dem *argumentum ad ignorantiam*. Walton (2016, S. 228) bezeichnet dies auch als „reasoning from absence of knowledge". Charakteristisch ist das Heranziehen eines Arguments für eine These, die Kenntnisse oder Wissen für diese Behauptung bewusst, d. h. absichtlich, oder unbewusst *ignoriert* (Walton, 2016). Es ist davon auszugehen, dass dieser Typ des Arguments hier überwiegend unbewusst eingesetzt wird, da die Probanden einen Fehlschluss aufgrund von mangelndem Wissen über Messunsicherheiten begehen. Es findet also kein bewusstes Ignorieren gegenläufiger Daten statt. Zudem liegt hier eine Situation vor, in der Schüler mit hoher Wahrscheinlichkeit einen Widerspruch zwischen den Daten und der eingangs aufgestellten Hypothese identifizieren. Bekanntermaßen neigen Lernende dazu, eher bestätigende Überprüfungen der eigenen Hypothese vorzunehmen. Dieses auch als *confirmation bias* bezeichnete Phänomen (Ganser & Hammann, 2009; Klay-

man & Ha, 1987; Nickerson, 1998; Zeidler, 1997) könnte insbesondere bei dieser Kategorie dazu führen, dass die Probanden hier vorliegende natürliche Streuungen in den Daten entsprechend ihrer zuvor aufgestellten Hypothese interpretieren und beispielsweise unbedeutende Mittelwertsunterschiede zur Stützung der eigenen falschen Eingangshypothese heranziehen.

MESSUNSICHERHEITEN (EXPLIZIT) Im Gegensatz dazu wurden in der Kategorie Messunsicherheiten (explizit) Aussagen erfasst, die auf eine bewusste Berücksichtigung von Unsicherheiten in den Daten schließen lassen. Aus dem Auftreten dieser Kategorie lässt sich folgern, dass Probanden die Unsicherheiten in Daten und deren Relevanz erkennen können, wie sich auch in anderen Arbeiten gezeigt hat (z. B. Lehrer, Kim & Schauble, 2007). Die Verwendung dieser Kategorie lässt aber nicht auf das Vorliegen von Kompetenzen im Umgang mit Messunsicherheiten schließen. Vielmehr handelt es sich um das Erkennen einer spezifischen Problematik, jedoch ohne die „Lösung" zu kennen. Offen bleibt hier, inwiefern die Verwendung dieser Argumentkategorie zu physikalisch korrekten Hypothesen führt. Zum einen ist denkbar, dass die Verwendung dieser Kategorie in einer Argumentation zum Wechseln oder Beibehalten einer physikalischen Hypothese nach dem Experiment dazu führt, dass Lernende eher eine physikalisch korrekte Hypothese annehmen, da die Auseinandersetzung mit Unsicherheiten in Messdaten Teil des naturwissenschaftlichen Erkenntnisprozesses ist (DeBoer, 2000; Heinicke, 2012; Hellwig, 2012; Roth, 2004). Es ist zum anderen denkbar, dass Lernende bei Verwendung dieser Kategorie Hypothesen eher konservativ behandeln. Im hier verwendeten Beispiel des Fadenpendels könnte das bedeuten, dass die Probanden auch nach dem Experiment eine Abhängigkeit zwischen Pendelmasse und Schwingungsdauer annehmen, da sie zwar Unsicherheiten erkennen und berücksichtigen, aber sich gegen die Annahme der Daten entscheiden, da sie beispielsweise die Variabilität der Daten nicht evaluieren können. Dieser Aspekt wird im weiteren Verlauf dieser Arbeit genauer untersucht (vgl. Forschungsfrage 2).

DATEN ALS EVIDENZ Daten gelten als ein strukturelles Element eines Arguments (vgl. Abschnitt 2.1). Es ist daher wenig verwunderlich, dass die Probanden in dieser Studie anhand von physikalischen Messdaten und Beobachtungen argumentieren. Dabei muss jedoch differenziert werden: Toulmin (2003) definiert Daten als die „Fakten", die als Grundlage für eine Behauptung herangezogen werden („facts we appeal to as the foundation for the claim", S. 90). Insbesondere im Kontext der vorliegenden Arbeit ist jedoch nicht sicher, dass es sich dabei immer um (naturwissenschaftliche) *Mess*daten handelt. Dies wird insbesondere im Kontrast z. B. zur Argumentkategorie Intuition

deutlich: Dort wird eine „Ahnung“ oder ein Bauchgefühl als „Fakt“ für eine Hypothese herangezogen. Die Kategorie Daten als Evidenz meint jedoch ganz explizit das Heranziehen von Messdaten als Evidenz. Daten und Evidenzen lassen sich anhand ihres epistemischen Status voneinander abgrenzen, wie Lederman et al. (2014) formulieren:

> Data and evidence serve different purposes in a scientific investigation. Data are observations gathered by the scientist during the course of the investigation, and they can take various forms (e. g., numbers, descriptions, photographs, audio, physical samples, etc.). Evidence, by contrast, is a product of data analysis procedures and subsequent interpretation, and is directly tied to a specific question and a related claim. (S. 70)

Vor diesem Hintergrund ist anzumerken, dass sämtlichen Aussagen, die dieser Kategorie zugeordnet wurden, unterstellt wird, zuvor eine Datenanalyse durchlaufen zu haben – auch wenn dies nicht explizit genannt wird, denn nur so gelten Messdaten nach Lederman et al. (2014) als Evidenz.

2.4 Übertragung des Elaboration-likelihood-model of persuasion auf das Argumentieren beim Experimentieren

Die im vorhergehenden Abschnitt dargestellte Kategorisierung differenziert die Typen der Argumente nach der *Art der Begründung* beim Beibehalten oder Wechseln einer physikalischen Hypothese (vgl. Abschnitt 2.1.3). Darüber hinaus ist es möglich, die Kategorien hinsichtlich der *Qualität* der zugrundeliegenden Informationsverarbeitung zu differenzieren. Dabei sind verschiedene Strategien denkbar. Im Folgenden soll ein Ansatz verfolgt werden, der auf einem Vorschlag von Gorsky und Finegold (1994) beruht. Gorsky und Finegold (1994) schlagen vor, Reaktionen von Schülern auf nicht-hypothesenkonforme Daten (vgl. Abschnitt 4.1.1) nach Rationalität bzw. Affektivität zu differenzieren. Dies ist insbesondere deshalb sinnvoll, da oftmals von der Annahme ausgegangen wird, dass ein Wechsel hin zu einem neuen naturwissenschaftlichen Konzept auf kognitiv-rationale Weise erfolgt. Diese unter dem Begriff *cold conceptual change* (Pintrich, Marx & Boyle, 1993; Sinatra, 2005) bekannte Perspektive vernachlässigt affektive und motivationale Faktoren sowie Merkmale der Lernsituation. Es hat sich aber gezeigt, dass auch andere, ebenfalls affektive Persönlichkeitsmerkmale wie Überzeugungen, Motivation oder das Interesse einen maßgeblichen Einfluss auf einen Konzeptwechsels haben (*hot conceptual change*). Vor diesem Hintergrund gibt es Bestrebungen, persuasive Prozesse beim Lehren und Lernen stärker zu untersuchen

(z. B. Alexander, Fives, Buehl & Mulhern, 2002; Chinn & Samarapungavan, 2001; Leach & Scott, 2003; Sinatra & Kardash, 2004), da Persuasion sowohl auf rationale als auch auf affektive Weise erreicht werden kann:

> When we persuade, we seek to change others' behaviors, or their understanding, judgments, or positions on a particular topic by appealing both to reason and emotion. (Alexander et al., 2002, S. 796)

Ein Ziel von Persuasion ist die Änderung einer Einstellung (Murphy & Mason, 2006). Unter dieser Definition wird die vorliegende Situation, d. h. hier das Experiment mit nicht-hypothesenkonformen Daten, als eine Überzeugungssituation verstanden, bei der Lernende unter Benutzung des Verstandes oder durch affektive Faktoren experimentelle Beobachtungen, Daten und Hinweisreize evaluieren und sich für oder gegen eine zuvor aufgestellte Hypothese entscheiden, d. h. ihre Einstellung gegenüber dem physikalischen Thema verändern. In diesem Prozess wird von den berichteten Argumentkategorien Gebrauch gemacht. Eine Klassifikation der Argumentkategorien nach dem Grad der Rationalität bzw. Affektivität erlaubt es daher, Rückschlüsse auf die zugrundeliegende Qualität der Informationsverarbeitung zu ziehen.

Dieser Ansatz weist Parallelen zu den sog. Zwei-Prozess-Modellen aus Kognitions-, Kommunikations- und Sozialpsychologie auf, welche die Informationsverarbeitung von Menschen in bestimmten Situationen beschreiben. Alle Modelle zeichnen sich dadurch aus, dass sie zwei grundlegende Qualitäten der Informationsverarbeitung unterscheiden, die sich – vereinfacht formuliert – als „rational" bzw. „affektiv" bezeichnen lassen. Diese Polarität lässt sich bei verschiedenen Modellen finden, z. B. *rational* und *experiential* in der Cognitive-Experiential-Self-Theorie (CEST) (Epstein, 1998; Epstein, Pacini, Denes-Raj & Heier, 1996; Pacini & Epstein, 1999), *systematisch* und *heuristisch* im Heuristisch-Systematischen Modell (HSM) (Eagly & Chaiken, 1993), *zentral* und *peripher* im Elaboration-Likelihood Model of Persuasion (ELM) (Petty & Cacioppo, 1986, 1996).

Insbesondere das ELM ist für die vorliegende Arbeit von besonderer Relevanz, da es detailliert Prozesse der Einstellungsänderung (*attitude change*, Petty & Cacioppo, 1986) durch persuasive Kommunikation beschreibt. Einstellungen sind dadurch charakterisiert, dass sie die Akzeptanz von Personen gegenüber Objekten ausdrücken, die nach einer sorgsamen Evaluation aller Gegebenheiten kognitiver und affektiver Art zum Ausdruck kommt („evaluative judgement", Maio & G. Haddock, 2009, S. 4). Nach dieser Definition sind hier eine Hypothese bzgl. eines physikalischen Zusammenhangs sowie die aus einem Experiment gewonnenen Daten und Beobachtungen als Objekte zu verstehen, denen gegenüber es nach einer Evaluation zu einer

Einstellungsänderung kommt (Annahme oder Ablehnen einer Hypothese bzw. Aufstellen einer neuen Hypothese). Da Einstellungen bzw. Überzeugungen (engl. *beliefs*) definitionsgemäß mit Konstrukten wie Wissen (engl. *knowledge*) überlappen, ist die Modellierung des Aufstellens von Hypothesen als Einstellungsänderung hier legitim (Murphy & Mason, 2006).

Das ELM unterscheidet bei der Evaluation von persuasiven Mitteilungen zwei mögliche Wege – die zentrale und die periphere Route. Verarbeitet der Empfänger einer persuasiven Mitteilung sie nach der zentralen Route, setzt er sich kognitiv mit den zentralen und sachrelevanten Inhalten der Nachricht auseinander. Dabei werden sämtliche zur Verfügung stehenden Informationen verarbeitet, indem er versucht, die Information zu verstehen, zu kombinieren und zu integrieren (Murphy & Mason, 2006; Petty & Cacioppo, 1996). Am Ende dieses Prozesses nimmt der Empfänger eine (idealerweise) kohärente Einstellung gegenüber dieser Nachricht ein. Petty und Cacioppo (1996, S. 256) fassen daher zusammen: "under the central route, persuasion is based on a thoughtful consideration of the object or issue at hand." Bei der Verarbeitung der Nachricht über die periphere Route sind andere Faktoren bedeutsam. Anstelle der Auseinandersetzung mit den Inhalten der Nachrichten werden periphere Hinweisreize („persuasion cues", Petty & Cacioppo, 1996, S. 256) in die Verarbeitung der Nachricht mit einbezogen. Diese Hinweisreize sind situative Faktoren, die es dem Empfänger der Nachricht ohne intensive kognitive Auseinandersetzung ermöglichen, die Nachricht zu evaluieren. Es werden z. B. die angenommene Kompetenz und Glaubwürdigkeit der Informationsquelle, die Art der Darstellung der Information, die Länge der Nachricht oder die reine Anzahl der Argumente betrachtet.

Die Verarbeitung einer persuasiven Nachricht über eine der beiden Routen hat Konsequenzen auf die Nachhaltigkeit der Einstellungsänderung. Es konnte gezeigt werden, dass die Verarbeitung einer persuasiven Mitteilung über die zentrale Route zu einer nachhaltigeren Einstellungsänderung führt, während die Verarbeitung über die periphere Route dazu führt, dass die Einstellungsänderung weniger stabil ist und damit eine geringere Nachhaltigkeit aufweist.

Das ELM stammt ursprünglich aus der Sozial- bzw. Kommunikationspsychologie. Hier soll es auf das Lernen von naturwissenschaftlichen Inhalten durch Experimentieren übertragen werden. Einen vergleichbaren Ansatz, der ebenfalls auf dem ELM beruht, haben Dole und Sinatra (1998) mit dem Cognitive Reconstruction of Knowledge Model (CRKM) vorgelegt. Das CRKM beschreibt Konzeptwechsel beim Lernen aus kognitiv-konstruktivistischer Perspektive und ist in weiten Teilen mit dem ELM vergleichbar. Das CRKM beschreibt ebenfalls ein sog. *engagement continuum*, motivationale Einflussgrößen auf personaler und situationaler Seite, sowie die Annahme, dass eine Verar-

beitung über eine zentrale und eine periphere Route erfolgen kann, wobei letztere zu eher schwachen und nicht dauerhaften Änderungen führt. Im Gegensatz zum ELM berücksichtigt das CRKM zudem die Stärke und das Engagement, mit dem Präkonzepte aufrecht erhalten werden, sowie soziale Effekte. Ein für die vorliegende Arbeit wesentlicher Faktor, nämlich die Fähigkeit zur Verarbeitung einer Nachricht wird jedoch im CRKM nicht berücksichtigt. Dieser Faktor birgt jedoch ein hohes didaktisches Potential (vgl. Abschnitt 2.6). Ferner modelliert das CRKM das Lernen durch Experimentieren nicht explizit als Persuasion. Aus diesen Gründen wird in der vorliegenden Arbeit das ELM dem CRKM vorgezogen.

2.4.1 *Der Einfluss personaler Faktoren auf die Verarbeitung*

Die Elaborationswahrscheinlichkeit wird nach Petty und Cacioppo (1986) auf der personalen Seite von zwei wesentlichen Faktoren beeinflusst: Zum einen müssen Menschen motiviert sein, zum anderen müssen sie kognitiv fähig und in der Lage sein („ability to process a message“, Petty & Cacioppo, 1986, S. 126), eine persuasive Nachricht zu verarbeiten.

Unter dem Begriff Motivation fassen Petty und Cacioppo (1986) verschiedene Konstrukte zusammen: Motivation wird zum einen beeinflusst durch die persönliche Relevanz – z. B. im Hinblick auf das zukünftige Leben –, die ein Thema für eine Person haben kann. Persönliche Relevanz zeichnet sich durch die intrisische Bedeutsamkeit eines Themas aus (Petty & Cacioppo, 1986). Erfährt eine Person eine persönliche Bedeutung in dem der persuasiven Mitteilung zugrundeliegenden Thema, beeinflusst dies die Motivation und damit die Wahrscheinlichkeit zur Elaboration.

Neben der persönlichen Relevanz beeinflusst das Kognitionsbedürfnis die Motivation: “Need for cognition refers to an indiviual’s tendency to engage in and enjoy effortful cognitive endeavors” (Cacioppo, Petty & Kao, 1984, S. 306). Das Konstrukt wurde in seiner gegenwärtigen Lesart von Cacioppo und Petty (1982) entwickelt, um interindividuelle Unterschiede in dem Bestreben von Personen zu beschreiben, sich intensiv und mit Freude anstrengenden und aufwendigen kognitiven Bemühungen auszusetzen. Dabei sehen Cacioppo und Petty (1982) das Kognitionsbedürfnis als relativ stabilen motivationalen *trait* an, es handelt sich also nicht um eine intellektuelle Fähigkeit (Cacioppo, Petty, Feinstein & Jarvis, 1996). Personen mit einem hohen Kognitionsbedürfnis besitzen dementsprechend eine hohe Motivation, sich mit anspruchsvollen Inhalten kognitiv auseinanderzusetzen. Demgegenüber stehen „cognitive misers“ (S. 197), d. h. Personen, die kognitiven Bemühungen eher auszuweichen bestrebt sind. In Bezug auf das ELM konstatieren Petty und Cacioppo (1986) auf der Grundlage verschiedener empirischer Ergebnisse: “If people high in need

for cognition tend to engage in and enjoy effortful cognitive activity, they should be particularly likely to evaluate a message by scrutinizing and elaborating the issue-relevant arguments presented" (S.151). Petty und Cacioppos beschreiben das Konstrukt „need for cognition" zunächst unabhängig von konkreten Situationen bzw. Inhalten. Es ist aber sehr plausibel, dass die Neigung zu einer intensiven und freudvollen kognitiven Auseinandersetzung mit komplexen Informationen stark von der Situation bzw. den Inhalten abhängig ist. Die Unspezifität des Kognitionsbedürnisses wird kritisiert (z. B. von Pechtl, 2009), ist aber von Cacioppo et al. (1996) auch selbst erkannt worden: "Also, situational factors can moderate cognitive motivation such that the motivation to think is so low that neither individuals low nor individuals high in need for cognition think about the material or is so high that both individuals low and high in need for cognition think extensively about the material" (S. 229).

Ein weiterer bestimmtender Faktor der Motivation ist das Interesse. Liegt ein gewisses Interesse an einer Situation vor, so erhöht dies ebenfalls die Elaborationswahrscheinlichkeit. Dies ist konform zu einer Vielzahl empirischer Arbeiten, die ebenfalls einen Zusammenhang zwischen dem (situationalen) Interesse und dem Lernen gezeigt haben (Dole & Sinatra, 1998; Hidi & Renninger, 2006).

Neben den motivationalen Komponenten beeinflusst die Fähigkeit, eine persuasive Nachricht zu verarbeiten und zu verstehen, die Elaborationswahrscheinlichkeit. Die zur Verarbeitung benötigte Fähigkeit hängt zum einen von Charakteristika der persuasiven Nachricht ab, z. B. der Verständlichkeit und der Schwierigkeit. Zum anderen benötigen Personen – je nach Schwierigkeit – ein hinreichendes Hintergrundwissen zum Verständnis der persuasiven Mitteilung. Ein höheres Hintergrundwissen erhöht daher die Elaborationswahrscheinlichkeit. Gleichzeitig findet aber auch eine Interaktion zwischen den motivationalen Komponenten und der Fähigkeit statt: Eine – z. B. aufgrund mangelnden Vorwissens – unverständliche Nachricht kann zu einer Frustation führen, was wiederum die Motivation senkt und zu einer geringeren Elaborationswahrscheinlichkeit führen kann, sodass die Nachricht eher über die periphere Route verarbeitet wird.

2.4.2 *Der Einfluss situationaler Faktoren auf die Verarbeitung*

Neben den aufgeführten personalen Faktoren konnten verschiedene empirische Untersuchungen zeigen, dass auch Charakteristika der Situation bzw. der persuasiven Mitteilung einen Einfluss auf die Elaborationswahrscheinlichkeit nehmen. Unter situativen Charakteristika können z. B. die Expertise, die Attraktiviät, Glaubwürdigkeit oder Sympathie der Quelle einer Nachricht oder die reine Anzahl der Argumente verstanden werden. Petty, Cacioppo und Goldman (1981) variierten die Elaborationswahrscheinlichkeit durch die Manipulation

der persönlichen Relevanz einer persuasiven Mitteilung und konnten zeigen, dass bei einer niedrigen persönlichen Relevanz bzw. einer damit einhergehenden niedrigen Elaborationswahrscheinlichkeit eher die Expertise der Quelle der Nachricht evaluiert wird. Nachrichten, die von einer Quelle höherer Expertise ausgingen, hatten dann eine stärker persuasive Wirkung auf die Rezipienten. In anderen Studien wurde die Fähigkeit zur Verarbeitung einer persuasiven Nachricht variiert. Dann zeigte sich ein ähnliches Bild: Probanden, die eine geringe Fähigkeit zur Elaboration aufwiesen, verließen sich eher auf Hinweisreize und evaluierten seltener sachrelevante Informationen. Bei Vorliegen einer hohen Elaborationswahrscheinlichkeit können Quellenfaktoren jedoch beim Auswerten der sachrelevanten Informationen hilfreich sein (Petty & Cacioppo, 1984).

2.4.3 *Klassifikation der Argumentkategorien nach Zentralität und Peripherität*

Werden die zentrale bzw. periphere Route aus dem ELM nun auf das dargestellte Kategoriensystem verschiedener Typen von Argumenten übertragen (vgl. Abschnitt 2.3), ergibt sich die folgende dichotome Aufteilung:

Die Argumente, die auf eine eher rationale Auseinandersetzung mit den aus dem Experiment gewonnenen Informationen schließen lassen, werden der zentralen Klasse zugeordnet. Hierzu zählen zum einen die beiden Argumentkategorien, die sich konkret auf die aus dem Experiment gewonnenen Messdaten beziehen, d. h. die Kategorie Daten als Evidenz bzw. die Kategorie Messunsicherheiten (explizit). Beide sind nach Petty und Cacioppo (1996) der zentralen Route zuzuordnen, weil es sich bei dem Auswerten der Messdaten sowie der Berücksichtigung von Messunsicherheiten um kognitive Aktivitäten handelt, die unmittelbar die Informationen der persuasiven Mitteilung (hier also das Experiment und die Ergebnisse daraus) zum Gegenstand haben. Zur zentralen Klasse zählen ferner die Kategorien Eignung des Experiments sowie Experimentierkompetenz (zentral), da hier eine sachlogische Auseinandersetzung mit dem experimentellen Aufbau sowie der eigenen experimentellen Fähigkeiten stattfindet. In der Auseinandersetzung mit dem Experiment zeugt die Verwendung dieser beiden Kategorien von der kognitiven Auseinandersetzung mit der Qualität der Messdaten bzw. experimentellen Beobachtungen.

In die periphere Klasse können die Argumentkategorien eingeordnet werden, die auf eine eher non-rationale Auseinandersetzung mit dem Experiment schließen lassen. Dazu zählt die Argumentkategorie Intuition, welche deutlich auf eine affektiv-intuitive Verarbeitung der aus dem Experiment erhaltenen Informationen hinweist. Bei der Verwendung der Argumentkategorie Expertenwissen haben sich Pro-

banden insbesondere auf die bereits in das Experiment bzw. in die Simulation implementierte Expertise berufen. Entsprechend dem ELM wird diese auf situativen und äußerlichen Hinweisreizen beruhende Verarbeitung hier der peripheren Klasse zugeordnet. Ferner können die Kategorien Ignoranz und Heranziehen einer (falschen) physikalischen Theorie der peripheren Klasse zugeordnet werden, da beide Kategorien davon zeugen, dass der Proband einer kognitiv-rationalen Auseinandersetzung mit den widersprüchlichen Messdaten ausweicht. Ebenfalls zur peripheren Klasse kann die Kategorie Experimentierkompetenz (peripher) gezählt werden, die auf subjektiven Eigenschaftszuschreibungen der eigenen Person im Sinne von Selbstwirksamkeitserwartungen beruht.

Die Kategorie Messunsicherheiten (implizit), die sich durch eine naive Interpretation der Variation in den Messdaten auszeichnet, wird hier ebenfalls in die periphere Klasse eingeordnet. Diese Einordnung ist aus fachphysikalischer Sicht legitim, denn eine tendenziöse Interpretation von Datenvariation deutet auf ein naives Bild über die Auswertung physikalischer Messungen sowie auf mangelndes Wissen über Messunsicherheiten hin. Aus Schülerperspektive betrachtet kann dieses Vorgehen jedoch auch als „rational“ bewertet werden, denn es ist anzunehmen, dass Mittelwerte bzw. einzelne Messungen auf Grundlage gelernter mathematischer Verfahren als „größer“, „kleiner“ bzw. „gleich“ bewertet werden. Diese Kategorie wäre aus dieser Perspektive dann eher der zentralen Klasse zuzuschreiben.

Die hier beschriebene Einteilung ist auch für die Kategorien Heranziehen einer (falschen) physikalischen Theorie sowie Expertenwissen unsicher. Es sind z. B. Situationen denkbar, in denen die Bewertung der bereits in das Experiment implementierten Expertise von Fachkollegen durchaus eher rationalen Kriterien folgt. In der Physik ist es insbesondere bei den aktuellen Großexperimenten (z. B. CERN, LIGO) sogar notwendig, die bereits in das Experiment implementierte Expertise zu berücksichtigen, da einzelne Forscher bzw. Forschergruppen den gesamten experimentellen Aufbau nicht alleine leisten können. Die vorgenommene Einteilung der Argumentkategorien stellt daher keine allgemein gültige Klassifikation dar, sondern muss stets im Kontext der Situation neu bewertet werden. Für den weiteren Verlauf der Arbeit ist die streitbare Zuordnung nicht weiter von Bedeutung, da zur weiterführenden Untersuchung nur relativ eindeutig zuzuordnende Kategorien ausgewählt wurden (Intuition, Daten als Evidenz, Messunsicherheiten (explizit) sowie Expertenwissen). Eine genauere Darstellung der Auswahlkriterien findet sich in Abschnitt 2.5.3. Tabelle 1 stellt die vorgenommene Einteilung der Argumentkategorien in einer Übersicht dar.

Die Verarbeitung einer persuasiven Mitteilung über eine der beiden beschriebenen Routen hat Konsequenzen auf die Qualität der erreichten Einstellungsänderung. So konnte gezeigt werden, dass die

Tabelle 1: Dichotomisierung der Argumentkategorien in eine periphere und eine zentrale Klasse

Periphere Klasse	Zentrale Klasse
Intuition	Experimentierkompetenz (zentral)
Ignoranz	Eignung des Experiments
Experimentierkompetenz (peripher)	Messunsicherheiten (explizit)
Heranziehen einer (falschen) physikalischen Theorie	Daten als Evidenz
Expertenwissen	
Messunsicherheiten (implizit)	

Verarbeitung einer persuasiven Mitteilung über die zentrale Route nachhaltigere Einstellungsänderungen hervorruft, während Einstellungsänderungen, die über die periphere Route erreicht werden, eher nur von kürzerer Dauer sind (Petty & Cacioppo, 1986, 1996).

Die hier beschriebene Unterteilung nach dem Grad der kognitiven Verarbeitung sachrelevanter Informationen in eine periphere bzw. zentrale Route beschreibt lediglich prototypische Prozesse an den Endpunkten eines zugrundeliegenden Kontinuums an Elaboration (Petty & Cacioppo, 1986), das sich von keinerlei bis hin zu einer intensiven kognitiven Verarbeitung sachrelevanter Informationen erstreckt. Elaboration wiederum beschreibt das Ausmaß der kognitiven Verarbeitung. In Bezug auf das „Beschreiten" der beiden Routen sind verschiedene Lesarten des ELM denkbar. In der vorliegenden Arbeit wird davon ausgegangen, dass das Argumentieren für oder gegen eine zuvor aufgestellte Hypothese sowohl über die periphere als auch gleichzeitig über die zentrale Route stattfinden kann. Diese Annahme steht in Widerspruch zu Naumann (2004), der einen „*trade-off* zwischen systematischen und heuristischen Prozessen der Einstellungsänderung annimmt: Es kommt (im Grundsatz) *entweder* zu einer Einstellungsänderung auf der peripheren *oder* zu einer Einstellungsänderung auf der zentralen Route" (S. 29, Hervorhebungen im Original). Die Annahme, dass beide Routen gleichzeitig beschritten werden können, ist aber plausibel, denn in den Interviews wurden relativ häufig sowohl Argumente der peripheren als auch der zentralen Klasse gleichzeitig in einer Begründung identifiziert (Post-hoc-Analysen ergaben, dass in den 81 Interviews mit den Probanden, die das Pendelexperiment durchgeführt haben, 55.5 % Probanden sowohl Argumentkategorien der zentralen als auch der peripheren Klasse verwendet haben).

2.5 AUSWAHL UND ANPASSUNG DER KONSTRUKTE AN DAS ARGUMENTIEREN BEIM EXPERIMENTIEREN

Ein für die vorliegende Studie zentraler Punkt ist die Übertragung wesentlicher Konstrukte und Wirkzusammenhänge, die sich aus dem ELM ergeben, auf das Lernen von Naturwissenschaften beim Experimentieren im Schulkontext. Das ELM nennt verschiedene Faktoren auf personeller Ebene, die Einfluss auf die Verarbeitungsqualität einer persuasiven Nachricht haben. Im folgenden Abschnitt sollen diese sehr unspezifischen Faktoren auf das Lernen von Physik übertragen werden. Es soll dabei der Frage nachgegangen werden, welche Entsprechungen die Einflussgrößen Fähigkeit, Interesse, Relevanz der Nachricht und Kognitionsbedürfnis beim Lernen von Naturwissenschaften haben.

In Abschnitt 2.3 wurde berichtet, dass Lernende beim Experimentieren zehn verschiedene Kategorien von Argumenten heranziehen. Lediglich vier dieser Kategorien finden im weiteren Verlauf der Arbeit Berücksichtigung. Es werden die Gründe für dieses Vorgehen dargelegt.

2.5.1 *Anpassung der motivationalen Konstrukte*

Das ELM nennt auf Personenebene drei motivationale Faktoren, welche die Verarbeitung über die periphere bzw. zentrale Route beeinflussen. Dabei ist zu berücksichtigen, dass diese verschiedene Konstrukte, nämlich das Kognitionsbedürfnis, die persönliche Relevanz einer Nachricht sowie das Interesse konzeptuell zusammengefasst und als „Motivation zur Verarbeitung sachrelevanter Informationen“ über die zentrale Route bezeichnet werden. Eine solche Herangehensweise widerspricht aktuelleren Konzeptionen, insbesondere zur Unterscheidung von Interesse und Motivation: „The decisive criterion of the interest construct, which enables it to be clearly distinguished from several neighbouring motivational concepts is its content specificity“ (Krapp & Prenzel, 2011, S. 30). Im Gegensatz zum ELM sollen daher in der vorliegenden Arbeit die drei motivationalen Konstrukte getrennt voneinander erhoben und ihr Einfluss untersucht werden.

Darüber hinaus erfordern die im ELM benannten Konstrukte Interesse und persönliche Relevanz eine differenzierte Betrachtung beim Übertragen auf das Lernen von Naturwissenschaften. Das Interesse wird in der aktuellen Forschung verstanden als eine „Person-Gegenstands-Beziehung, welche die psychischen Phänomene des Lernens und der Entwicklung als (permanente) Austauschbeziehung zwischen einer Person und ihrer sozialen und gegenständlichen Umwelt interpretiert“ (Krapp, 2001, S. 286). Zentral ist dabei die Beziehung einer Person zu einem Gegenstand, wobei der Gegenstand hier sowohl den thematischen Bereich (Fadenpendel, Mechanik, Physik) als auch

die Tätigkeit (Aufstellen einer Hypothese, Durchführung des Experiments, Evaluation von Daten und Beobachtungen) meint (Krapp, 2001). Bei Vorliegen von Interesse ist dieser Gegenstand für die Person von einer emotionalen bzw. wertbezogenen Valenz geprägt. In der Folge kommt es dann zu einer Bereitschaft zur Auseinandersetzung. Die Entwicklung des Interesses kann gemäß aktueller Konzeptionen in verschiedene Phasen differenziert werden (Hidi & Renninger, 2006; Mitchell, 1993; Renninger & Hidi, 2011). Das *individuelle* Interesse beschreibt dabei relativ persistente Strukturen, während das *situationale* Interesse durch die „aktuellen Anregungsbedingungen" (Krapp, 2001, S. 287) bzw. der Interessantheit eines Gegenstandes in einer spezifischen Situation definiert ist. In der vorliegenden Arbeit wird ein Einfluss des situationalen Interesses auf die Verwendung bestimmter Argumentkategorien angenommen, da dieses sowohl den Merkmalen der Experimentiersituation als auch der wahrgenommenen Interessantheit des Gegenstands Rechnung trägt. Es konnte in verschiedenen Studien der positive Einfluss des situationalen Interesses auf die Qualität der Informationsverarbeitung gezeigt werden: Es ist bekannt, dass das situationale Interesse die Aufmerksamkeit, das Erreichen von Zielsetzungen und den Lernerfolg positiv beeinflusst (für einen Überblick siehe z. B. Hidi & Renninger, 2006).

Bei der Übertragung der theoretischen Überlegungen und Befunde zum ELM müssen zudem die unterschiedlichen Kontexte und Rahmenbedingungen der Informationsrezeption beachtet werden. Während das ELM aus kommunikationspsychologischen Fragestellungen heraus entwickelt wurde, ist das vorliegende Forschungsvorhaben im Kontext der schulischen Physikunterrichts angesiedelt. Im unterrichtlichen Geschehen liegen jedoch (mehr oder weniger) klare Zielsetzungen für die Informationsverarbeitung vor. Es kommt daher nicht zwingend zu einer individuell geprägten Informationsselektion. Dies ist insbesondere für das Konstrukt der persönlichen Relevanz von Bedeutung. Es ist anzunehmen, dass die persönliche Relevanz im Physikunterricht eine andere Rolle als im ELM hat, da die Auseinandersetzung mit dem Unterrichtsgegenstand durch die Lehrkraft vorgegeben ist und nicht „freiwillig" erfolgt. Es ist zudem nicht davon auszugehen, dass die wahrgenommene persönliche Relevanz einer einzelnen Experimentiersituation zum Thema Fadenpendel besonders stark ausgeprägt ist, da die spezifische Thematik des Experiments für die Lernenden von eher geringer Bedeutung ist. Hinsichtlich des genannten Einflusses der persönlichen Relevanz einer Nachricht wird daher hier das allgemeinere Konstrukt der Werteinschätzung der Naturwissenschaft herangezogen. Dies ist relevant, da insbesondere „[n]euere Studien (Hulleman, Godes, Hendricks & Harackiewicz, 2010; Hulleman & Harackiewicz, 2009) [...] deutlich auf die zentrale Bedeutung der Werteinschätzung für die Entwicklung von Interesse, Studien-

wahl und Leistung, speziell auch im Kontext naturwissenschaftlichen Schulunterrichts“ (Knogler & Lewalter, 2014, S. 5) hinweisen.

2.5.2 *Die Fähigkeit zum Verarbeiten von Messdaten und experimentellen Beobachtungen*

Zur Verarbeitung einer persuasiven Nachricht über die zentrale Route müssen Personen nicht nur hinreichend motiviert sein, sondern auch die Fähigkeit zur Verarbeitung besitzen („ability to process issue-relevant arguments“, Petty & Cacioppo, 1986, S. 130). Übertragen auf das physikalische Experimentieren sind dabei verschiedene Entsprechungen denkbar. Zum einen könnten Fähigkeiten zum Umgang mit Messdaten und Messunsicherheiten eine Rolle spielen. Verfahren zu ihrer Erfassung liegen jedoch bisher noch nicht vor bzw. werden zurzeit erst entwickelt (z. B. Schulz & Priemer, 2016). Zum anderen ist ein Einfluss des themenspezifischen Vorwissens zu erwarten. Das themenspezifische Vorwissen umfasst hier nicht nur das verwendete Experiment (Fadenpendel, vgl. Abschnitt 4.1), sondern das gesamte domänenspezifische Wissen aus dem Inhaltsbereich Mechanik. Aus diesem Grund wurde in der vorliegenden Arbeit die Fähigkeit zur Verarbeitung einer Nachricht aus dem ELM durch das Fachwissen Mechanik ersetzt, denn inhaltsspezifische Vorwissenseffekte beim Experimentieren sind vielfach berichtet worden, z. B. hinsichtlich des Umgangs mit Hypothesen, der Anwendung von Strategiewissen und der Variablenkontrolle (z. B. Künsting et al., 2008). Zudem existieren empirische Belege, die darauf hindeuten, dass das Fachwissen im Bereich Mechanik als ein guter Prädiktor für physikalisches Wissen (Friege & Lind, 2004), für das physikalische Problemlösen (Brandenburger, 2017) oder für die Elektrizitätslehre gelten kann (Cappell, 2013, die letztgenannten Arbeiten auf Ebene von Lehramtsstudierenden).

2.5.3 *Auswahl der untersuchten Argumentkategorien*

Das in Abschnitt 2.3 hergeleitete Kategoriensystem enthält zehn Argumentkategorien. Im weiteren Verlauf der Arbeit sollen diese Argumentkategorien genutzt werden, um den Umgang von Lernenden mit aus einem Experiment gewonnenen Informationen (z. B. Messdaten, Beobachtungen) zu charakterisieren. Da kein geeignetes Verfahren zur Erfassung der Stärke der Verwendung der Argumentkategorien vorlag, musste ein solches Verfahren im Rahmen dieser Arbeit zunächst entwickelt werden (zur Operationalisierung der Argumentkategorien siehe Abschnitt 4.3.1, die Testentwicklung ist in Anhang B dargestellt). Da die Testentwicklung im Rahmen dieser Arbeit nur für einen Teil der Argumentkategorien auf einem entsprechenden Niveau umgesetzt werden konnte, wurden für den weiteren Verlauf

der Studie vier Kategorien mit aus wissenschaftlicher und schulischer Perspektive hoher didaktischer Relevanz ausgewählt.

Aus der zentralen Klasse wurden die Kategorien Daten als Evidenz und Messunsicherheiten (explizit) ausgewählt. Zum einen ist das Heranziehen von Messdaten als Evidenz sowie ein adäquater Umgang mit Messunsicherheiten wesentlicher Bestandteil des Experimentierens (Heinicke, 2012; Hellwig, 2012; Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, 2004). Beide Kategorien zeichnen sich daher durch eine hohe didaktische Relevanz aus. Im Sinne des ELM handelt es sich bei beiden Kategorien um eine für die Verarbeitung über die zentrale Route repräsentative Kategorie, die auf eine rationale und sachlogische Auseinandersetzung mit den Experimentierergebnissen schließen lässt.

Aus der peripheren Klasse wurden die Kategorien Intuition und Expertenwissen ausgewählt. Bei der Kategorie Intuition handelt es sich in Sinne des ELM um eine Argumentkategorie, die stark auf eine Verarbeitung über die periphere Route hindeutet. Des weiteren sind intuitive Prozesse im naturwissenschaftlichen Unterricht wenig erforscht, aber von hoher Relevanz (Fensham & Marton, 1992). In dieser Arbeit soll u. a. der Vergleichbarkeit der Verarbeitung von Informationen aus Real- bzw. Computerexperimenten nachgegangen werden (Forschungsfrage 3). Vor diesem Hintergrund weist die Kategorie Expertenwissen eine besondere Relevanz auf. Es ist zu vermuten, dass Probanden, die mit dem Simulationsexperiment arbeiten, sich stärker auf das in einer Simulation implementierte Expertenwissen berufen, da sie berücksichtigen, dass den Programmen ein mathematisches Modell zugrundeliegt, das von einem Experten entwickelt und implementiert wurde. Schulisches Lernen ist zudem stark von der Expertise anderer abhängig. Die Gestaltung und Auswahl von Materialien, Schulbüchern, Experimenten wird von Personen vorgenommen, die eine hohe Expertise (fachlich und didaktisch) aufweisen. Der Rückgriff auf das Expertenwissen bei der Argumentation für oder gegen eine Hypothese ist daher besonders relevant.

Bei der Auswahl der Argumentkategorien spielten zudem pragmatische Gründe eine Rolle. So wurde berücksichtigt, dass die Forschungsfragen (vgl. Abschnitt 3.2) eine quantitative Methodik zur Erfassung der Stärke der Verwendung der Argumentkategorien erfordern. Die quantitative Operationalisierung einiger Kategorien erscheint jedoch problematisch (z. B. für die Argumentkategorie Ignoranz).

Damit fanden die Kategorien Daten als Evidenz, Intuition, Messunsicherheiten (explizit) sowie die Kategorie Expertenwissen Eingang in den weiteren Verlauf der Arbeit. Die übrigen Argumentkategorien wurden in weitere Untersuchungen bisher nicht eingeschlossen.

2.6 INTEGRATION VON THEORIE UND VORARBEITEN UND ABLEITUNG DES FORSCHUNGSINTERESSES

In den vorangegangenen Abschnitten wurden der theoretische Hintergrund zum Argumentieren im Physikunterricht (vgl. Abschnitt 2.1), die Rolle von Computersimulationen im Physikunterricht und in der naturwissenschaftsdidaktischen Forschung (vgl. Abschnitt 2.2), die eigenen Vorarbeiten zur Kategorisierung von Argumentem beim Experimentieren sowie die sich daraus ergebende Analogie zum ELM dargestellt (vgl. Abschnitt 2.3). In diesem Abschnitt wird der referierte Stand der Forschung mit den eigenen Vorarbeiten zusammengeführt, um daraus das Forschungsinteresse der vorliegenden Arbeit abzuleiten.

In der vorliegenden Arbeit werden Experimentiersituationen verwendet, bei denen Lernende auf nicht-hypothesenkonforme Daten treffen. Die Evaluation dieser Daten und Beobachtungen geschieht in einem Argumentationsprozess, bei dem unter Rückgriff auf verschiedene Kategorien von Argumenten (vgl. Abschnitt 2.3) eine zuvor aufgestellte Hypothese beibehalten bzw. verworfen wird. Die aufgeführte Analogie zum ELM betrachtet diese Situation als die Verarbeitung einer Überzeugungsbotschaft, deren Qualität (zentral vs. peripher) durch eine Reihe von motivationalen Faktoren, die Fähigkeit des Empfängers sowie situationale Faktoren beeinflusst wird. Dabei soll die Verarbeitung einer persuasiven Nachricht über die zentrale Route zu einer nachhaltigeren Einstellungsänderung führen. Die aus den Vorarbeiten abgeleiteten Argumentkategorien wurden hinsichtlich ihres zentralen bzw. peripheren Charakters diskutiert und die aus dem ELM genannten Faktoren an das Experimentieren angepasst. Abbildung 1 zeigt eine Zusammenfassung des ELMs, übertragen auf die hier vorliegende Situation des naturwissenschaftlichen Experimentierens.

Bisher liegen kaum empirische Studien vor, die eine derartige Analogie zwischen dem Lernen von Naturwissenschaften durch Experimentieren und persuasiven Entscheidungsprozessen ziehen. Die Prüfung der Anwendbarkeit des ELM beim Lernen von Naturwissenschaften durch Experimentieren hat eine hohe didaktische Relevanz: Für den Unterricht ist es wünschenswert zu wissen, welche Faktoren das Beschreiten der zentralen Route beim Umgang mit Hypothesen und experimentellen Daten begünstigen, denn nach dem ELM führt eine Verarbeitung über die zentrale Route zu einer nachhaltigeren Einstellungsänderung, hier in Form des Wechsels von einer fachlich inkorrekten zu einer (idealerweise) korrekten Hypothese. Es ist daher zu prüfen, ob und in welchem Ausmaß die aus dem ELM genannten Faktoren eine Rolle bei der Verarbeitung von experimentellen Daten und Beobachtungen spielen und welchen Einfluss die Verwendung einzelner Argumentkategorien auf die Nachhaltigkeit dieser Entschei-

dung nehmen. Obwohl dies im ELM nicht explizit benannt ist, interessiert hier auch die Qualität der Entscheidung, d. h. die Richtigkeit der aufgestellten Hypothese. Ferner ist das Beschreiten der zentralen Route bei der Verarbeitung der aus einem physikalischen Experiment gewonnenen Informationen der aus physikdidaktischer Perspektive wünschenswertere Weg der Verarbeitung: Ein Ziel naturwissenschaftlicher Bildung ist es, Schüler zu befähigen, Ansichten und Aussagen zu naturwissenschaftlichen Themen evidenzbasiert zu begründen (vgl. Kapitel 1). Dies korrespondiert klar mit der zentralen Route der Verarbeitung, da Messdaten und Beobachtungen als Evidenzen herangezogen, Unsicherheiten in Daten berücksichtigt und der experimentelle Aufbau sowie die Durchführung des Experiments in die Evaluation einbezogen werden (vgl. Tabelle 1).

In Abschnitt 2.2 wurde dargelegt, dass die aktuelle Forschungslage zurzeit keine eindeutige Aussage darüber erlaubt, an welcher Stelle und in welchem Ausmaß durch die Art des Mediums – hier ein reales Experiment bzw. eine Computersimulation – induzierte Unterschiede in den Experimentierprozessen und den Prozessen kognitiver Verarbeitung auftreten. Es wurde darüber hinaus gezeigt, dass bisherige Vergleiche auf völlig unterschiedlichen abhängigen Variablen beruhen, z. B. Lernleistung, Lerneffizienz, die Problemlöseleistung, Planen von Experimenten oder Experimentierkompetenz.

Das ELM nennt als einflussnehmende Faktoren auch Charakteristika der Situation. Übertragen auf die vorliegende Arbeit ist daher anzunehmen, dass Lernende allein aufgrund der Unterschiede zwischen Real- bzw. Computerexperimenten Informationen aus den beiden Settings unterschiedlich verarbeiten: Zum einen ist es plausibel, dass Probanden eine Entscheidung zum Beibehalten oder Verwerfen eher unter Rückgriff auf die Argumentkategorie Expertenwissen bestreiten, da eine Computersimulation bereits eine durch einen Experten in das Programm implementierte analytische Lösung eines physikalischen Problems enthält. Zum anderen muss geprüft werden, ob die Möglichkeit, in Simulationen sehr viel schneller als in Realexperimenten viele Daten zu erzeugen, nicht eine komplexe Informationsflut hervorruft, die sich nicht mehr sachgerecht verarbeiten lässt. Probanden könnten in einer solchen Situation eine Hypothese eher unter Rückgriff auf Argumente der Kategorie Intuition begründen. In diesem Fall können solche Heuristiken dazu führen, dass evidente Informationen unbeachtet bleiben und Lernende eher bei ihren Präkonzepten bleiben. Diese beiden Hypothesen würden bedeuten, dass Probanden, die an einem Computerexperiment arbeiten, eher die periphere Route der Informationsverarbeitung bestreiten.

Es sind aber auch Einflusswege seitens des situationalen Faktors „Medium“ auf die Argumentkategorien der zentralen Route zu überprüfen: Aus Computerexperimenten reproduzierte Messwerte weisen –unter sonst identischen Bedingungen – eine beliebig hohe Präzision

auf. Es ist daher anzunehmen, dass Probanden, die mit Computerexperimenten arbeiten, weniger die Relevanz von Messunsicherheiten erkennen und diese Argumentkategorie entsprechend seltener in ihrer Argumentation verwenden. Streng genommen können aus epistemologischer Perspektive Daten, die aus Computerexperimenten reproduziert werden, nicht als naturwissenschaftliche empirische Evidenz bezeichnet werden. Es ist denkbar, dass Lernende diesen epistemischen Unterschied erkennen und für Daten aus Computerexperimenten eine geringere Evidenzzuschreibung vornehmen. Diese Annahme kann anhand der Stärke der Verwendung der Argumentkategorie Daten als Evidenz zwischen den Gruppen geprüft werden.

Es geht hierbei nicht per se um eine grundsätzliche Einschätzung eines Mediums hinsichtlich der Frage, welches „besser" als das andere ist. Aus physikdidaktischer Perspektive ist diese Frage auch nicht zielführend, da Entscheidungen zum Einsatz von Computersimulationen im Physikunterricht weniger auf dem Bestreben nach einer „optimalen" Experimentiersituation basieren, sondern insbesondere abhängig von Zielsetzungen, aber auch der Verfügbarkeit oder aus Gründen der Ökonomie getroffen werden (Hofstein & Lunetta, 2004). Wünschenswert ist aber Kenntnis darüber, welches Medium welche kognitiven Prozesse begünstigt. Durch die Analyse von Argumentationen zum Umgangs mit eigenen Hypothesen soll in der vorliegenden Arbeit daher ein Ansatz verfolgt werden, der genau diese Frage beantwortet.

Die grundlegende Frage ist dabei, welche Unterschiede es zwischen realen und virtuellen Experimenten in der Argumentation von Lernenden für das Beibehalten bzw. Verwerfen einer Hypothese gibt. Dieser Ansatz ist hoch relevant, denn er beantwortet für die Praxis, inwieweit reale Experimente im Physikunterricht durch Computersimulationen ersetzt werden können. Zum anderen trägt dieser Ansatz dazu bei, die Generalisierbarkeit von Befunden zum Experimentieren besser zu beurteilen, die rein durch den Einsatz von Computersimulationen gewonnen wurden (vgl. Abschnitt 2.2).

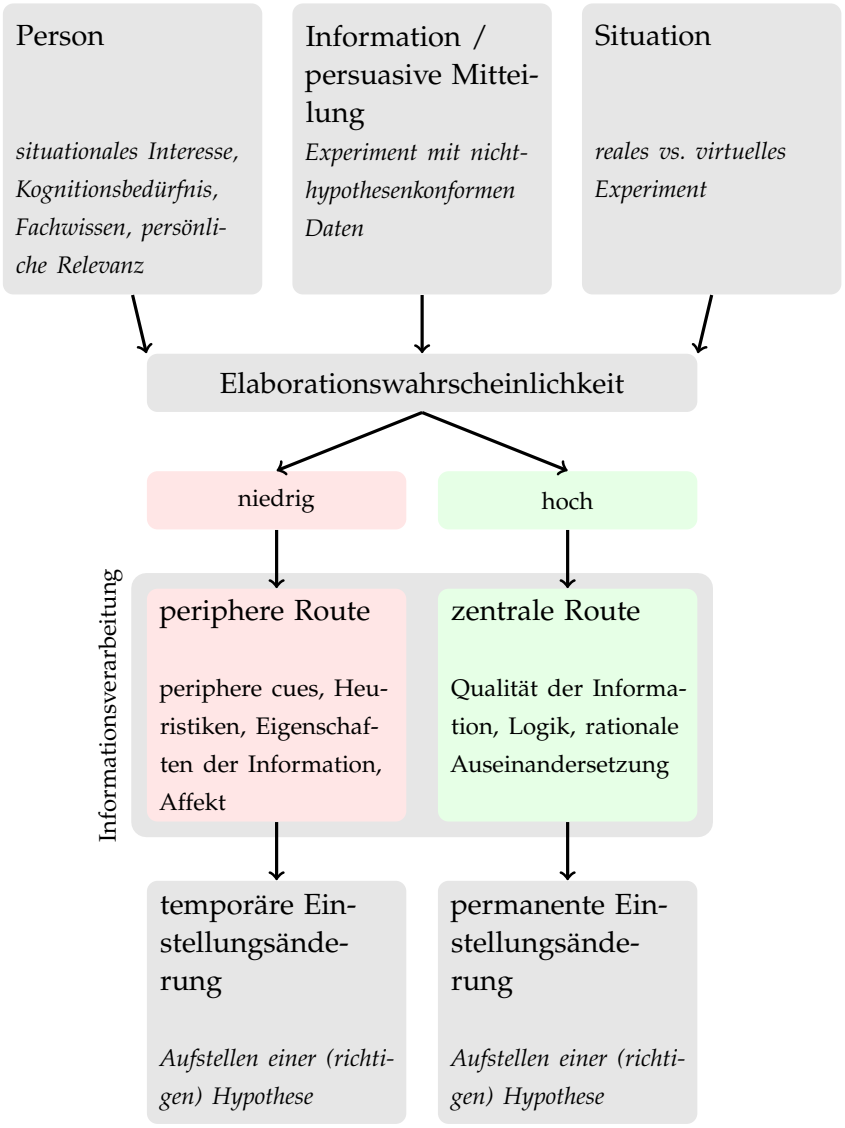


Abbildung 1: Übertragung des ELMs nach Petty und Cacioppo auf das naturwissenschaftliche Experimentieren. Vorgenommene Anpassungen an den Kontext der vorliegenden Arbeit sind kursiv gedruckt.

3 FRAGESTELLUNG

In diesem Kapitel werden zunächst die Ziele der Untersuchung zusammengefasst. Im folgenden Abschnitt werden die Forschungsfragen dargelegt und die statistischen Hypothesen aufgestellt.

3.1 ZIELE DER UNTERSUCHUNG

Die vorliegende Arbeit verfolgt zwei Hauptziele. Zum einen soll überprüft werden, ob die aus den Vorarbeiten angenommene Analogie zwischen dem ELM und den von Schülern vorgebrachten Argumenten beim Wechseln oder Beibehalten einer Hypothese vor dem Hintergrund selbstständig generierter Daten auch empirisch gestützt werden kann. Dazu werden der Einfluss verschiedener Persönlichkeitsmerkmale auf die unterschiedlichen Argumentkategorien beim Experimentieren und deren Einfluss wiederum auf die Richtigkeit gewählter Hypothesen und die Nachhaltigkeit dieser Entscheidung untersucht (Forschungsfragen 1 und 2). Damit wird erstmals geprüft, ob das ELM oder Teile davon für das Lernen von Naturwissenschaften von Bedeutung sein können. Aus didaktischer Perspektive relevant ist dabei insbesondere die Frage, ob und unter welchen Voraussetzungen nachhaltiges Lernen stattfindet und ob dies der kognitiven Verarbeitung auf der zentralen Route entspricht.

Zum anderen soll empirisch geprüft werden, ob sich Argumentationen für oder gegen eine gewählte Hypothese und die Nachhaltigkeit dieser Entscheidung beim Experimentieren mit realen Experimenten bzw. Computersimulationen unterscheiden (Forschungsfrage 3). So lässt sich exemplarisch einschätzen, ob Lernende anhand experimentell gewonnener Daten und Beobachtungen in beiden Settings zu vergleichbaren naturwissenschaftliche Ergebnisinterpretationen kommen. Aus Perspektive schulischen Physikunterrichts liefert die Untersuchung damit Hinweise auf die Gestaltung von Unterricht mit Experimenten. Zudem kann exemplarisch eingeschätzt werden, ob Forschungsergebnisse zum naturwissenschaftlichen Experimentieren, die ausschließlich auf der Basis von Computersimulationen gewonnen wurden, auf das Experimentieren mit realen Materialien übertragbar sind. Die Vergleichbarkeit des Arbeitens in realen und virtuellen Experimenten ist eine grundlegende Frage, denn mit einer zunehmenden Nutzung von virtuellen anstelle von realen Experimenten in Lernprozessen geht ein substanzieller methodischer Wechsel einher, der – zumindest aus didaktischer und erkenntnistheoretischer Sicht – andere naturwissenschaftliche Ergebnisinterpretationen nach sich zieht.

3.2 FORSCHUNGSFRAGEN UND STATISTISCHE HYPOTHESEN

Bei den statistischen Hypothesen handelt es sich um gerichtete Alternativhypothesen (Forschungshypothesen), die dem theoretisch begründbaren Effekt entsprechen (Döring & Bortz, 2016, S. 53). Die Nullhypothesen sind der Übersichtlichkeit halber nicht notiert, negieren aber den jeweiligen postulierten Effekt oder behaupten das Gegenteil.

In Bezug zum ELM ergeben sich die folgenden Forschungsfragen (FF):

FF 1: Lassen sich die aus dem ELM theoretisch vorhergesagten Einflüsse von fachlichem Vorwissen, situationalem Interesse, persönlicher Relevanz und Kognitionsbedürfnis auf die Stärke der Verwendung bestimmter Argumentkategorien (Daten als Evidenz, Intuition, Messunsicherheiten (explizit) und Expertenwissen) beim Wechseln bzw. Beibehalten einer Hypothese empirisch nachweisen?

Es wird vermutet, dass das fachliche Vorwissen, das situationale Interesse, die persönliche Relevanz und das Kognitionsbedürfnis signifikante Beiträge zur Vorhersage der Stärke der Verwendung der vier verschiedenen Argumentkategorien liefert: Es wird ein positiver Zusammenhang dieser Variablen bzgl. der Kategorien Daten als Evidenz und Messunsicherheiten (explizit) und ein negativer bzgl. der Kategorien Intuition und Expertenwissen (periphere Route) erwartet. Vor diesem Hintergrund lassen sich 16 statistische Hypothesen formulieren.

$H_{1.1}$: Je niedriger das fachliche Vorwissen, desto mehr argumentieren die Probanden mit der Kategorie Intuition.

$H_{1.2}$: Je niedriger das fachliche Vorwissen, desto mehr argumentieren die Probanden mit der Kategorie Expertenwissen.

$H_{1.3}$: Je höher das fachliche Vorwissen, desto mehr argumentieren die Probanden mit der Kategorie Messunsicherheiten (explizit).

$H_{1.4}$: Je höher das fachliche Vorwissen, desto mehr argumentieren die Probanden mit der Kategorie Daten als Evidenz.

$H_{1.5}$: Je niedriger das Kognitionsbedürfnis, desto mehr argumentieren die Probanden mit der Kategorie Intuition.

$H_{1.6}$: Je niedriger das Kognitionsbedürfnis, desto mehr argumentieren die Probanden mit der Kategorie Expertenwissen.

$H_{1.7}$: Je höher das Kognitionsbedürfnis, desto mehr argumentieren die Probanden mit der Kategorie Messunsicherheiten (explizit).

$H_{1.8}$: Je höher das Kognitionsbedürfnis, desto mehr argumentieren die Probanden mit der Kategorie Daten als Evidenz.

$H_{1.9}$: Je niedriger die persönliche Relevanz, desto mehr argumentieren die Probanden mit der Kategorie Intuition.

$H_{1.10}$: Je niedriger die persönliche Relevanz, desto mehr argumentieren die Probanden mit der Kategorie Expertenwissen.

$H_{1.11}$: Je höher die persönliche Relevanz, desto mehr argumentieren die Probanden mit der Kategorie Messunsicherheiten (explizit).

$H_{1.12}$: Je höher die persönliche Relevanz, desto mehr argumentieren die Probanden mit der Kategorie Daten als Evidenz.

$H_{1.13}$: Je niedriger das situationale Interesse, desto mehr argumentieren die Probanden mit der Kategorie Intuition.

$H_{1.14}$: Je niedriger das situationale Interesse, desto mehr argumentieren die Probanden mit der Kategorie Expertenwissen.

$H_{1.15}$: Je höher das situationale Interesse, desto mehr argumentieren die Probanden mit der Kategorie Messunsicherheiten (explizit).

$H_{1.16}$: Je höher das situationale Interesse, desto mehr argumentieren die Probanden mit der Kategorie Daten als Evidenz.

FF 2: Lassen sich die aus dem ELM theoretisch vorhergesagten Einflüsse der Zentralität bzw. Peripherität der Argumentkategorien auf die Entscheidung zum Wechseln bzw. Beibehalten einer Hypothese und hinsichtlich der Nachhaltigkeit dieser Entscheidung empirisch nachweisen?

Es wird angenommen, dass Probanden, die mehr mit den Argumentkategorien der zentralen Route argumentieren, eher von einer falschen Hypothese zu einer richtigen Hypothese wechseln und bei dieser Entscheidung eine hohe Nachhaltigkeit bzgl. der nach dem Experimentieren angenommenen Hypothese zeigen. Zudem wird angenommen, dass Probanden, die eher mit Argumentkategorien der peripheren Route arbeiten, seltener von einer falschen Hypothese zu einer richtigen wechseln und bei dieser Entscheidung eine geringere Nachhaltigkeit zeigen. Daraus lassen sich die folgenden statistischen Hypothesen formulieren:

$H_{2.1}$: Je mehr die Probanden mit der Argumentkategorie Intuition argumentieren, desto seltener wechseln sie von einer falschen auf eine fachlich adäquate Hypothese.

$H_{2.2}$: Je mehr die Probanden mit der Argumentkategorie Expertenwissen argumentieren, desto seltener wechseln sie von einer falschen auf eine fachlich adäquate Hypothese.

$H_{2.3}$: Je mehr die Probanden mit der ArgumentKategorie Messunsicherheiten (explizit) argumentieren, desto eher wechseln sie von einer falschen auf eine fachlich adäquate Hypothese.

$H_{2.4}$: Je mehr die Probanden mit der Argumentkategorie Evidenz argumentieren, desto eher wechseln sie von einer falschen auf eine fachlich adäquate Hypothese.

Und bezüglich der Nachhaltigkeit des Hypothesenwechsels:

$H_{2.5}$: Je mehr die Probanden mit der Argumentkategorie Intuition argumentieren, desto niedriger ist die Nachhaltigkeit des Hypothesenwechsels.

$H_{2.6}$: Je mehr die Probanden mit der Argumentkategorie Expertenwissen argumentieren, desto niedriger ist die Nachhaltigkeit des Hypothesenwechsels.

$H_{2.7}$: Je mehr die Probanden mit der ArgumentKategorie Messunsicherheiten (explizit) argumentieren, desto höher ist die Nachhaltigkeit des Hypothesenwechsels.

$H_{2.8}$: Je mehr die Probanden mit der Argumentkategorie Evidenz argumentieren, desto höher ist die Nachhaltigkeit des Hypothesenwechsels.

Bezüglich der Unterschiede in der Verwendung bestimmter Argumente zwischen Probanden, die mit Realexperimenten bzw. mit Computersimulationen arbeiten, lassen sich die folgenden Forschungsfragen und statistischen Hypothesen herleiten:

FF 3: Welchen Einfluss hat der Typ der Experimentierumgebung – real oder virtuell – auf die Stärke der Verwendung bestimmter Argumentkategorien beim Wechseln bzw. Beibehalten einer Hypothese?

Es wird vermutet, dass Probanden, die in virtuellen Lernumgebungen arbeiten, geringer mit den Argumentkategorien Daten als Evidenz und Messunsicherheiten (explizit) sowie stärker mit den Kategorien Intuition und Expertenwissen argumentieren als Probanden, die in realen Lernumgebungen arbeiten. Daraus lassen sich die folgenden statistischen Hypothesen formulieren:

$H_{3.1}$: Probanden, die mit dem Computerexperiment gearbeitet haben, zeigen bzgl. des Hypothesenwechsels eine stärkere Verwendung der Argumentkategorie Intuition als die Probanden, die mit dem Realexperiment gearbeitet haben:

$$\overline{\mu}_{\text{int, R}} < \overline{\mu}_{\text{int, V}}$$

$H_{3.2}$: Probanden, die mit dem Computerexperiment gearbeitet haben, zeigen bzgl. des Hypothesenwechsels eine stärkere Verwendung der Argumentkategorie Expertenwissen als die Probanden, die mit dem Realexperiment gearbeitet haben:

$$\overline{\mu}_{\text{exp, R}} < \overline{\mu}_{\text{exp, V}}$$

$H_{3.3}$: Probanden, die mit dem Computerexperiment gearbeitet haben, zeigen bzgl. des Hypothesenwechsels eine geringere Verwendung der ArgumentKategorie Messunsicherheiten (explizit) (explizit) als die Probanden, die mit dem Realexperiment gearbeitet haben:

$$\overline{\mu}_{\text{mu, R}} > \overline{\mu}_{\text{mu, V}}$$

$H_{3.4}$: Probanden, die mit dem Computerexperiment gearbeitet haben, zeigen bzgl. des Hypothesenwechsels eine geringere Verwendung der Argumentkategorie Daten als Evidenz als die Probanden, die mit dem Realexperiment gearbeitet haben:

$$\overline{\mu}_{\text{evi, R}} > \overline{\mu}_{\text{evi, V}}$$

Die Forschungsfragen sind in Abbildung 2 schematisch zusammengefasst.

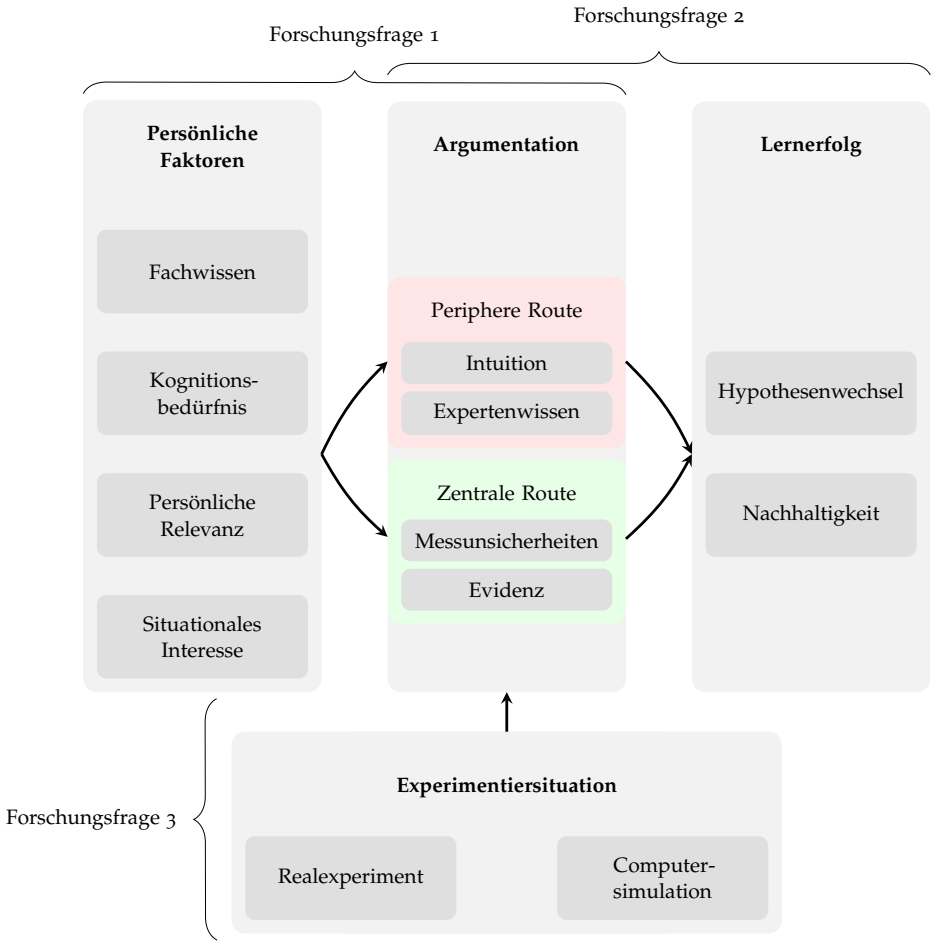


Abbildung 2: Schema der Forschungsfragen. Die Pfeile stellen untersuchte Ursachen- und Wirkzusammenhänge dar. Forschungsfrage 2 untersucht den Einfluss der Verwendung bestimmer Argumentkategorien auf die Entscheidung zum Hypothesenwechsel und dessen Nachhaltigkeit. Forschungsfrage 1 untersucht den Einfluss der persönlichen Faktoren auf die Verwendungen bestimmter Argumentkategorien. Forschungsfrage 3 untersucht den Einfluss der Experimentiersituation auf die Argumentation.

4

METHODEN

4.1 BESCHREIBUNG DER EXPERIMENTIERSITUATIONEN

4.1.1 *Kriterien für die Auswahl der Experimentiersituation*

Wie auch in den Vorarbeiten (vgl. Abschnitt 2.3) wurde zur Untersuchung der aufgeführten Fragestellungen ein Experiment zum Fadenpendel ausgewählt. Der Auswahl lagen die folgenden Kriterien zugrunde:

1. Die Konfrontion mit nicht-hypothesenkonformen Daten kann als ein Auslöser für eine intensive Argumentation gesehen werden (vgl. Abschnitt 2.1.5). Daher soll das ausgewählte physikalische Thema bei einer großen Zahl von Probanden eine Fehlvorstellung adressieren, sodass die Probanden vor dem Experiment eine fachlich inadäquate Hypothese aufstellen.

2. Die fachinhaltlich falsche Hypothese soll sich von Schülern der 8. und 9. Jahrgangsstufen eigenständig und experimentell überprüfen lassen (vgl. Abschnitt 4.2) und keine hohen Anforderungen an experimentelle Fähigkeiten und das Strategiewissen stellen. So soll gewährleistet werden, dass die Probanden gleichermaßen in der Lage sind, eigenständig ihre selbst aufgestellte Hypothese zu überprüfen. Diese Einschränkung ist hier legitim, da die vorliegende Arbeit sich auf die Prozesse des Erhebens von experimentellen Daten, der Auswertung sowie der Argumentation anhand der Daten und experimentellen Beobachtungen beschränkt.

3. Die Experimentiersituation muss sich sowohl im Real- als auch im Simulationsexperiment leicht und ökonomisch abbilden lassen. Dabei sollen die experimentellen Möglichkeiten in beiden Umgebungen möglichst gleich sein und sich außer den medienbedingten Differenzen keine weiteren (physikalischen) Unterschiede ergeben (vgl. Abschnitt 2.2).

4. Die Experimentiersituation soll schulrelevant sein und sich in den Lehrplänen wiederfinden.

Das Fadenpendel erfüllt alle unter Punkt 1 bis Punkt 4 genannten Voraussetzungen: Es ist bekannt, dass die Schwingungsdauer von der überwiegenden Zahl der Schüler als von der Masse abhängig betrachtet wird. Dies ist, mit den üblichen Einschränkungen im Schulkontext,

falsch (vgl. Abschnitt 4.1.2). In einer Arbeit von Kanari und Millar (2004) betrachteten 90 % der Probanden die Schwingungsdauer als von der Masse abhängig (vgl. Decker, 2012). Der Anspruch der experimentellen Durchführung ist als gering einzustufen. Diesem Aspekt wurde zudem dadurch Rechnung getragen, dass die Fadenlänge fixiert wurde, so dass diese Variable im Laufe des Experiments nicht variiert werden konnte (siehe Abschnitt 4.1.4).

Das Thema ist ferner als schulrelevant zu bewerten, wie hier exemplarisch anhand der Lehrpläne zweier Bundesländer gezeigt wird: Im nordrhein-westfälischen Lehrplan für das Fach Physik in der Sekundarstufe II ist das Pendel durch den Sachbereich „Mechanik“ mit den Themenfeldern „Harmonische Schwingung“, „Schwingungsvorgänge und Schwingungsgrößen“ sowie „Vorhersagbarkeit des Schwingungsverhaltens“ abgedeckt. In der Sekundarstufe I kann das Fadenpendel im Inhaltsfeld „Kraft, Druck, mechanische und innere Energie“ exemplarisch eingesetzt werden (Ministerium für Schule und Weiterbildung, Wissenschaft und Forschung des Landes Nordrhein-Westfalen, 1999, 2008). Im Rahmenlehrplan für das Land Berlin findet sich das Fadenpendel im Themenfeld „Schwingungen und Wellen“ in den Jahrgangsstufen 9 und 10 (Senatsverwaltung für Bildung, Jugend und Sport, 2006).

4.1.2 *Eine theoretische Betrachtung der Physik des Fadenpendels*

Bei der physikalischen Beschreibung des realen Fadenpendels werden in der Schule üblicherweise eine Reihe von Modellannahmen getroffen, die unter dem Begriff „mathematisches Pendel“ bekannt sind. Zu diesen Modellannahmen gehören u. a. die Vernachlässigung von Form und Ausdehnung der angehangenen Masse und Faden sowie des Auftriebs in Luft, von Auslenkungen $> 5°$, der Elastizität des Pendelfadens, des Luftwiderstands usw. (für eine genaue Darstellung aller Modellannahmen siehe Nelson & Olsson, 1986). Im folgenden Abschnitt soll zunächst eine mögliche theoretische Modellierung des mathematischen Fadenpendels dargelegt werden. Dabei wird diskutiert, zu welcher Abweichung die Kleinwinkelnäherung führt. Dies ist relevant, da davon auszugehen ist, dass die Schüler bei der Durchführung des Experiments in der vorliegenden Arbeit nicht immer nur im Bereich der Gültigkeit der Kleinwinkelnäherung arbeiten. Darauf aufbauend wird das Modell um Form und Ausdehnung der angehangenen Masse ergänzt, so dass eine Abschätzung vorgenommen werden kann, inwiefern die verwendeten Experimentiermaterialien, die natürlich eine Verletzung der Annahmen des mathematischen Pendels darstellen, geeignet sind, den Zusammenhang zwischen Masse und Schwingungsdauer sinnvoll zu untersuchen.

DAS MATHEMATISCHE PENDEL Das mathematische Fadenpendel kann durch eine Masse m, die an einem Faden konstanter Länge l aufgehangen ist, beschrieben werden. Nach Auslenkung um den Winkel φ vollzieht diese Masse eine Bewegung auf einem Kreisbogen mit dem Radius l um einen Aufhängepunkt. Wenngleich sich das Fadenpendel dazu eignet, die Äquivalenz von schwerer und träger Masse zu demonstrieren, soll hier nicht zwischen beiden unterschieden werden (Nolting, 2011).

In der einfachsten physikalischen Beschreibung des Fadenpendels gelten zunächst zwei Annahmen: Der Faden sei masselos und die Masse in einem Punkt konzentriert. Auf die Punktmasse wirken zwei Kräfte, die Fadenkraft (oder auch Fadenspannung) $\vec{F}_F$, die den Pendelkörper am freien Fall hindert, sowie die Gewichtskraft $\vec{F}_g$. Die gesamte auf den Pendelkörper wirkenden Kraft ergibt sich aus der Vektorsumme $\vec{G}$ dieser beiden Kräfte. Für die Änderung des Geschwindigkeitsbetrages des Pendelkörpers ist die tangentiale Komponente der Gesamtkraft $\vec{G}$ verantwortlich, die hier der Tangentialkomponente F_φ der Gewichtskraft $\vec{F}_g$ entspricht. Die Kräfte am Fadenpendel sind in Abbildung 3 dargestellt. Die Abbildung folgt dem Vorschlag von Backhaus (2000), bei der Darstellung der Kräfteverhältnisse am Fadenpendel die Situation bei $\dot{\varphi} \neq 0$ zu zeigen, da die in Lehrbüchern (z. B. Tipler & Mosca, 2006, S. 440) und Schulbüchern (z. B. Grehn, 1991, S. 111) oftmals aufzufindende Darstellung bei $\dot{\varphi} = 0$ nur einen Sonderfall zeigt, in dem $|\vec{F}_F| = F_r$ und die resultierende Kraft in Bewegungsrichtung wirkt, d. h. in tangentiale Richtung. Diese Darstellung stützt das bei Lernenden oftmals vorherrschende Präkonzept, dass die Kraft in Richtung der Bewegung und nicht in Richtung der Beschleunigung wirkt (Backhaus, 2000). Bei $\dot{\varphi} \neq 0$ zeigt die Gesamtkraft, die auf den Pendelkörper wirkt, jedoch stets nach innen!

Bei der Verwendung ebener Polarkoordinaten ergibt sich die Gewichtskraft zu

$$\vec{F}_g = F_r \vec{e}_r + F_\varphi \vec{e}_\varphi \; . \tag{1}$$

Unter Berücksichtigung der Trigonometrie lassen sich die Komponenten daher als

$$F_r = mg \cos\varphi \text{ und} \tag{2}$$

$$F_\varphi = -mg \sin\varphi \tag{3}$$

darstellen. Mit der Beschleunigung in ebenen Polarkoordinaten, die sich aus der zweifachen Differenziation des Ortsvektors zu

$$\ddot{\vec{r}} = \left(\ddot{r} - r\dot{\varphi}^2\right) \vec{e}_r + \left(2\dot{r}\dot{\varphi} + r\ddot{\varphi}\right) \vec{e}_\varphi \tag{4}$$

ergibt, kann die allgemeine Bewegungsgleichung für das mathematische Pendel erhalten werden:

$$m \left[\left(\ddot{r} - r\dot{\varphi}^2\right) \vec{e}_r + \left(2\dot{r}\dot{\varphi} + r\ddot{\varphi}\right) \vec{e}_\varphi\right] = \left(F_r + F_F\right) \vec{e}_r + F_\varphi \vec{e}_\varphi \; . \tag{5}$$

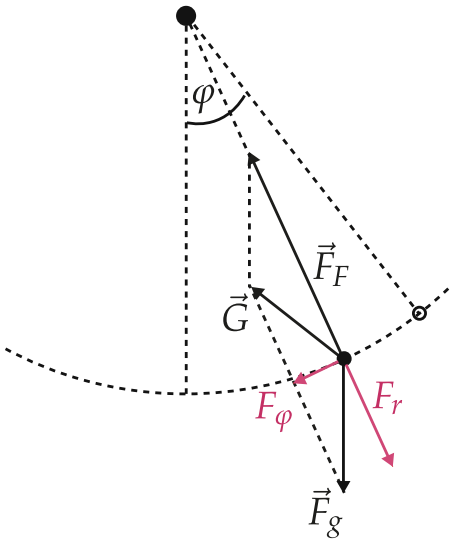


Abbildung 3: Kräfte und Definition des Koordinatensystems am Fadenpendel. Dargestellt ist das Kräfteverhältnis bei $\dot{\varphi} \neq 0$.

Bei der eingangs erwähnten Fadenkraft bzw. Fadenspannung $\vec{F}_F$ handelt es sich um eine Zwangskraft, die den konstanten Abstand vom Aufhänge- bzw. Drehpunkt als Zwangsbedingung sicherstellt:

$$r = l = \text{const. bzw. } \dot{r} = \ddot{r} = 0 \,. \tag{6}$$

Mit dieser Annahme vereinfacht sich Gl. 5 unter Zuhilfenahme von Gl. 2 in radialer Richtung zu einem eindimensionalen Problem:

$$F_F = -mg\cos\varphi - ml\dot{\varphi}^2 \tag{7}$$

An dieser Stelle wird deutlich, dass die von Backhaus (2000) kritisierte Darstellung mit $|\vec{F}_F| = F_r$ nur für den stationären Anfangsfall ($\dot{\varphi} = 0$) gültig ist.

Die Bewegung in $\vec{e}_\varphi$ - Richtung ergibt sich aus den Gleichungen 3, 5 und 6 zu

$$ml\ddot{\varphi} = -mg\sin\varphi \text{ bzw.} \tag{8}$$

$$\ddot{\varphi} + \frac{g}{l}\sin\varphi = 0 \,. \tag{9}$$

Gl. 9 beschreibt die Bewegung des mathematischen Pendels in tangentialer Richtung. Die Lösung dieser nichtlinearen Differenzialgleichung zweiter Ordnung führt auf elliptische Integrale 1.Art, die keine analytische Lösung zulassen. Mit der Kleinwinkelnäherung $\sin\varphi \approx \varphi$ für kleine Auslenkungen ergibt sich die harmonische Schwingungsgleichung

$$\ddot{\varphi} + \frac{g}{l}\varphi = 0 \text{ bzw. mit } \omega^2 = \frac{g}{l} \tag{10}$$

$$\ddot{\varphi} + \omega^2\varphi = 0 \,, \tag{11}$$

mit der Kreisfrequenz ω. Gl. 11 lässt sich durch den Ansatz $\varphi(t) = \varphi_0 \cos(\omega t + \delta)$ lösen; dabei ist φ_0 die Auslenkung bei $t = 0$ und δ die Phasenverschiebung. Die Schwingungsdauer T, also die Zeit, die das Pendel für eine volle Schwingung von 2π benötigt, ergibt sich zu

$$T = \frac{2\pi}{\omega} \text{ , bzw. mit Gl. 10 zu} \tag{12}$$

$$T = 2\pi\sqrt{\frac{l}{g}} \, . \tag{13}$$

Die Schwingungsdauer des mathematischen Pendels ist bei kleinen Auslenkungen wegen Gl. 13 sowohl von der Masse als auch von der Auslenkung φ_0 bei $t = 0$ unabhängig (Nolting, 2011; Scheck, 2007; Tipler & Mosca, 2006).

Bei Auslenkung des Pendels um Winkel $> 5\circ$ gilt Gl. 13 nicht mehr uneingeschränkt, die Auslenkung φ_0 nimmt dann Einfluss. Dies kann in der approximativen Lösung des elliptischen Integrals (basierend auf Gl. 9) durch Entwicklung in eine Reihe gezeigt werden (Nelson & Olsson, 1986):

$$T^* = 2\pi\sqrt{\frac{l}{g}}\left(1 + \frac{1}{16}\varphi_0^2 + \frac{11}{3072}\varphi_0^4 + \frac{173}{737280}\varphi_0^6 + \cdots\right) \, . \tag{14}$$

Selbst bei einer Berücksichtigung des 5. Gliedes der Reihenentwicklung beträgt bei einer Auslenkung von $\varphi_0 = 5^\circ$ die Abweichung $1 - T/T^*$ nur 0,048 %, bei $\varphi_0 = 45^\circ$ liegt die Abweichung bei 3,84 %. Der Einfluss der Auslenkung ist folglich so gering, dass er im Folgenden vernachlässigt werden kann. Im Experiment kann der Einfluss durch Konstanthalten von φ_0 eliminiert werden.

DAS PHYSIKALISCHE PENDEL Wie eingangs dargelegt, ist weiterhin die Abschätzung der Abweichung zwischen dem Modell des mathematischen Pendels und dem Modell des physikalischen Pendel von Bedeutung. Das physikalische Pendel berücksichtigt, dass sich die Masse des Pendels nicht in einem Punkt konzentriert; damit gehen Form und Ausdehnung der Pendelmasse ein. Da die Schüler in der vorliegenden Arbeit mit einem Aufbau experimentieren, bei dem sich der Massenmittelpunkt beim Anhängen weiterer Massestücke in Richtung des Aufhängepunkts verschiebt (einen Überblick über die Experimentiersituationen liefert Abschnitt 4.1.4), wird dieser Aspekt in der folgenden theoretischen Betrachtung berücksichtigt.

Unter Berücksichtigung von Form und Ausdehnung der angehangenen Masse M des Fadenpendels in Form des Trägheitsmomentes I, ergibt sich für die Schwingungsdauer T_{phys} des physikalischen Pendels unter Berücksichtigung der Kleinwinkelnäherung:

$$T_{\text{phys}} = 2\pi\sqrt{\frac{I}{Mgl}} \, . \tag{15}$$

Das physikalische Pendel berücksichtigt also – anders als das mathematische Pendel –, dass die Masse des Pendels nicht in einem Punkt am unteren Ende des Fadens konzentriert ist.

Das Trägheitsmoment I_{Zyl} eines Zylinders, der um eine Achse senkrecht zur Zylinderachse im Schwerpunkt rotiert, ergibt sich zu

$$I_{\text{Zyl}} = \frac{1}{12}Mh^2 + \frac{1}{4}Mr^2 \; ; \tag{16}$$

dabei gibt h die Länge und r den Radius an (Lipschutz, Spiegel & Liu, 2009, S. 41). Da die Rotationsachse um die Fadenlänge l_{F} verschoben ist, kann mittels des Steiner'schen Satzes das Trägheitsmoment zu

$$I_{\text{Pendel}} = I_{\text{Zyl}} + Ml_F^2 = \frac{1}{12}Mh^2 + \frac{1}{4}Mr^2 + Ml_F^2 \tag{17}$$

bestimmt werden. M steht hier für die Gesamtmasse von n angehangenen Schlitzgewichten der Masse m, $M = nm$. Durch Einsetzen von Gl. 17 in Gl. 15 ergibt sich ein Ausdruck für die Schwingungsdauer für eine zylinderförmige Masse, die an einem Faden der Länge l_F aufgehangen ist:

$$T_{\text{phys}} = 2\pi\sqrt{\frac{\frac{1}{12}h^2 + \frac{1}{4}r^2 + l_F^2}{gl_F}} \tag{18}$$

An Gl. 18 ist ersichtlich, dass die Schwingungsdauer auch im Modell des physikalischen Pendel von der Masse unabhängig ist (eine homogene Massenverteilung vorausgesetzt).

In der hier verwendeten Experimentiersituation werden zylinderförmige Schlitzgewichte verwendet, die auf einen Gewichtsteller aufgesteckt werden. Daher muss die Fadenlänge l_{F} durch die effektive Fadenlänge ersetzt werden. Die effektive Fadenlänge l_{eff} verkürzt sich bei n aufgesteckten Schlitzgewichten um die halbe Länge des aus n Schlitzgewichten der Höhe h bestehenden Zylinders:

$$l_{\text{eff}} = l_F - \frac{n}{2}h \; . \tag{19}$$

Unter Berücksichtigung von Gl. 19 ergibt sich aus Gl. 18 ein Ausdruck für die Schwingungsdauer $T_{\text{phys, eff}}$, der das Aufstecken von Schlitzgewichten auf den Gewichtsteller und die damit verbundene Änderung des Trägheitsmoments und die aufgrund des gewählten experimentellen Aufbaus bedingte Änderung der effektiven Fadenlänge berücksichtigt:

$$T_{\text{phys, eff}} = 2\pi\sqrt{\frac{\frac{1}{12}h^2 + \frac{1}{4}r^2 + (l_f - \frac{n}{2}h)^2}{g(l_F - \frac{n}{2}h)}} \tag{20}$$

Anhand der dargestellten Ausdrücke für die verschiedenen Modelle wurden nun für $n = 1 \ldots 3$ Massestücke die zu erwartenden

Tabelle 2: Theoretisch berechnete Schwingungsdauern für das physikalische Pendel (T_{phys}), das physikalische Pendel unter Berücksichtigung der sich ändernden effektiven Fadenlänge ($T_{\text{phys,eff}}$) und das mathematische Pendel unter Berücksichtigung der sich ändernden effektiven Fadenlänge (T_{eff}), exemplarisch für $n = 1 \ldots 3$ Massestücke.

n	M/kg	T_{phys}/s	$T_{\text{phys,eff}}/\text{s}$	T_{eff}/s
1	0,05	1,7944	1,763	1,763
2	0,10	1,7944	1,731	1,730
3	0,15	1,7944	1,698	1,697

Schwingungsdauern berechnet. Die verwendeten Zahlenwerte orientierten sich dabei an den Maßen des experimentellen Aufbaus (d. h. $h = 1{,}4$ cm, $r = 1{,}4$ cm, $l_{\text{F}} = 0{,}8$ m). Die Ergebnisse sind in Tabelle 2 dargestellt.

Der in der vorliegenden Arbeit eingesetzte experimentelle Aufbau, bei dem die effektive Fadenlänge durch das Auflegen von zusätzlichen Massestücken verkürzt wird, führt dazu, dass die Schwingungsdauern sich um etwa drei- bis siebenhundertstel Sekunden verkürzen (siehe Spalte $T_{\text{phys,eff}}$ in Tabelle 2). Dieser Effekt kann mit den zur Verfügung gestellten Materialien von den Schüler nicht aufgedeckt werden, da die Unsicherheitsquellen die dafür notwendige Präzision der Messung überdecken (zur Diskussion von Messunsicherheit und Präzision des hier verwendeten Aufbaus vgl. Abschnitt 4.1.4).

4.1.3 *Ein schultaugliche Herleitung eines Ausdrucks für die Schwingungsdauer*

Die oben vorgestellte Herleitung des Zusammenhangs zwischen Dauer der Schwingung und Pendelmasse ist aufgrund der Differenzialgleichung nicht unbedingt für den Einsatz in der Sekundarstufe geeignet. Daher soll hier eine weitere Möglichkeit der Herleitung des Zusammenhangs kurz angerissen werden, die für den Einsatz in der Schule geeignet ist. Wird die Bewegung eines sog. Kreispendels (konischen Pendels) untersucht und sind die Gesetzmäßigkeiten der Kreisbewegungen und die Ausdrücke für Zentripetal- bzw. Zentrifugalkraft bekannt, lässt sich eine Gleichung für die auf die Masse wirkenden Kräfte aufstellen, aus der sich unmittelbar der gesuchte Zusammenhang ergibt. Um zu zeigen, dass dieser Ausdruck nicht nur für das Kreispendel, sondern auch für das ebene Fadenpendel gilt, kann die Projektion des Schattenwurfs eines Kreispendels und eines ebenen Pendels, die beide mit gleicher Amplitude schwingen, auf eine

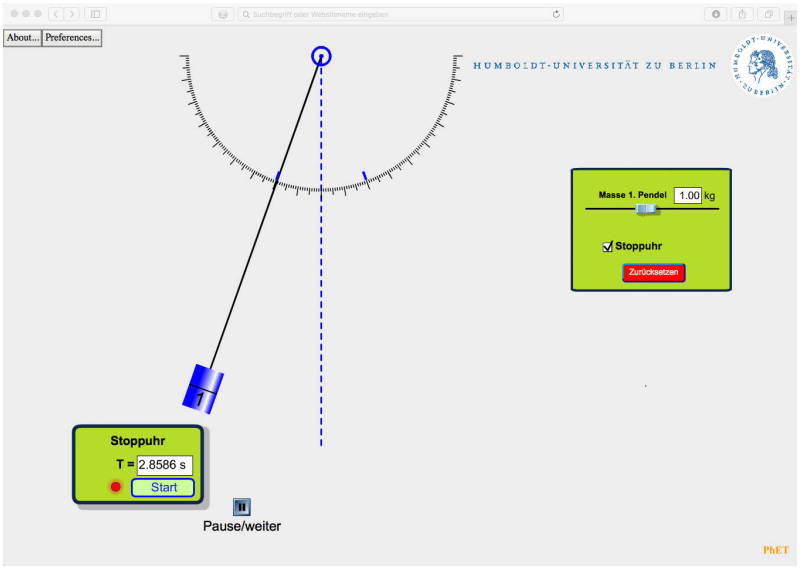


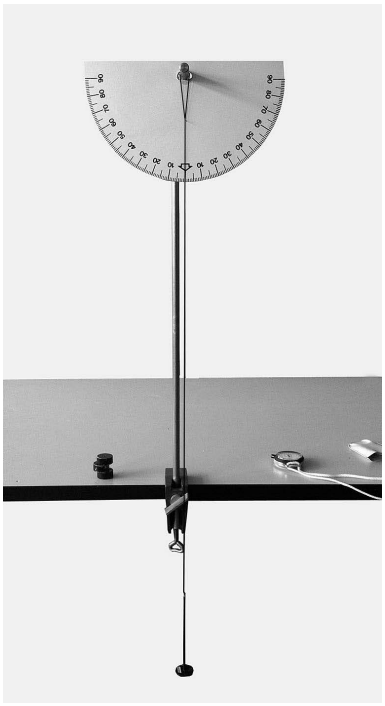
Abbildung 4: Das Simulationsexperiment. Das Kontrollfeld auf der rechten Seite erlaubt die Variation der Masse. Die Stoppuhr wird über das Feld unten links gesteuert.

Wand herangezogen werden (für eine umfassende Darstellung siehe z. B. Lüders & von Oppen, 2012, S. 133).

4.1.4 *Gegenüberstellung des realen und virtuellen Experiments zum Fadenpendel*

Um Probleme während des Experimentierens zu vermeiden, wurde in der Entwicklung der Experimente der Experimentierraum (Klahr & Dunbar, 1988) in seiner Bandbreite eingeschränkt. Da nur der Zusammenhang zwischen Pendelmasse und Schwingungsdauer, nicht aber der Zusammenhang zwischen Fadenlänge und Schwingungsdauer untersucht werden soll, wurde die Möglichkeit zur Variation der Variable „Fadenlänge“ nicht geboten. Die Fadenlänge wurde sowohl im realen als auch im virtuellen Experiment als konstant vorgegeben. Wie in Abschnitt 4.1.2 gezeigt, hat die Auslenkung keinen Einfluss auf die Schwingungsdauer. Da diese Variable aber nur schwerlich aus dem Experimentraum zu entfernen ist (andernfalls hätte ein konkreter Winkel vorgeben müssen) und da sich in der Voruntersuchung gezeigt hat, dass Probanden die Auslenkung intuitiv kontrollieren wollen, wird auf einer Winkelscheibe die Kontrolle der Auslenkung ermöglicht.

Das Realexperiment wurde mit üblichen Materialien aufgebaut (u.ä. Schlitzgewichte und Gewichtsteller der Fa. Phywe). Der Aufbau ist in Abbildung 5 dargestellt. Zur Variation der Pendelmasse liegen drei Massestückchen der Masse $m = 50\,\mathrm{g}$ bereit, die auf den passenden Halter aufgesteckt werden können. So stehen insgesamt vier Variationsmöglichkeiten zur Verfügung (Gewichtsteller plus drei Masse-



(a)

(b)

(c)

Abbildung 5: Das Realexperiment: (a) Kompletter Aufbau, Fadenlänge rund 80 cm; (b) Verwendete digitale Stoppuhr; (c) Gewichtsteller und Schlitzgewichte (zum Einfluss vom Trägheitsmoment und Änderung der effektiven Fadenlänge siehe Abschnitt 4.1.2).

stücke). Es sei hier angemerkt, dass durch das Aufstecken der Massestücke auf den Gewichtsteller ein systematischer Fehler vorliegt, da sich die effektive Fadenlänge so verkürzt. Es wurde aber in Abschnitt 4.1.2 gezeigt, dass die so verursachten Abweichungen vernachlässigbar klein sind und aufgrund der Messunsicherheit nicht aufgedeckt werden können. Zum Stoppen der Zeit steht eine digitale Stoppuhr zur Verfügung, die mit einer Hand bedient werden kann. Die Auslenkung kann an der Aufhängung abgelesen werden, wobei Parallaxenfehler berücksichtigt werden müssen. Das Ausmessen der Schwingungsdauer ist auf verschiedenen Wegen möglich, es bietet sich an, mit einer Hand das Pendel auszulenken und mit der anderen Hand die Stoppuhr zu bedienen. Dabei können ein oder mehrere Perioden gemessen werden; im letzteren Fall muss die gemessene Zeit durch die Anzahl der Perioden dividiert werden.

Das verwendete Simulationsexperiment basiert auf der Software *pendulum-lab* des Projekts PhET Interactive Simulations (2011). Die Simulation ist in den Programmiersprachen Adobe Flash und Actionsscript 3 geschrieben und kann mit dem Adobe Flash Player einfach und plattformunabhängig in einem Webbrowser dargestellt werden. Da der Quellcode der Simulation unter der GNU General Public License v2.0 (Free Software Foundation [FSF], 2011) zur Verfügung steht, konnten Änderungen am Programm vorgenommen werden. Das Programm wurde den Anforderungen entsprechend angepasst. Die für die vorliegende Arbeit verwendete Version der Simulation lässt lediglich die Manipulation der Masse, der Auslenkung und das Ausmessen der Schwingungsdauer mit einer digitalen Stoppuhr zu. Ferner wurden die Farben angepasst und ein Logo der Humboldt-Universität eingebettet. Die Simulation ist in Abbildung 4 dargestellt.

Bei der Entwicklung von Realexperiment und Computersimulation wurde darauf geachtet, dass die experimentellen Handlungsmöglichkeiten in beiden Settings identisch sind. Dennoch ergeben sich naturgemäß Differenzen zwischen beiden Settings, die hier im Folgenden dargestellt werden sollen.

- Die Masse des Pendel kann im Realexperiment durch das Anhängen der drei zur Verfügung gestellten Massestücke in drei Stufen variiert werden (vier verschiedene Messungen sind möglich, wenn eine Messung nur mit Gewichtsteller durchgeführt wird). Im Simulationsexperiment kann die Pendelmasse stufenlos über einen Schieberegler eingestellt werden.

- Die Schwingungsdauer wird im Realexperiment mit einer digitalen Stoppuhr erfasst. Dazu müssen eine bzw. mehrere Perioden mit einer hinreichend genauen Betätigung der Stoppuhr vermessen werden. Die Simulation bietet die Möglichkeit, die Schwingungsdauer mit einer digitalen Stoppuhr zu erfassen,

die automatisch von einem Nulldurchgang bis zum übernächsten Nulldurchgang die Zeit für eine Periode „misst“.

- In der Konsequenz ergibt sich aus den experimentellen Unterschieden auch ein Unterschied in der Unsicherheit der Messungen: Das Simulationsexperiment berechnet die Schwingungsdauer exakt auf drei signifikante Stellen. Die Standardmessunsicherheit ist daher null, und entsprechend ist die Mess- und Vergleichspräzision unendlich hoch (Hellwig, 2012; Kirkup & Frenkel, 2006). Aufgrund der manuellen Betätigung der Stoppuhr und der damit einhergehenden, durch die Reaktionslatenz bedingten Messunsicherheit kann die Schwingungsdauer im Realexperiment hingegen nur innerhalb des Überdeckungsintervalls von $T = \mu_T \pm 0,5$ s gemessen werden, wobei μ_T den Bestwert bezeichnet (da der menschliche Faktor die größte Ursache der Messunsicherheit ist, werden weitere Einflussgrößen nicht diskutiert).

- Im Simulationsexperiment kann das Pendel zunächst ausgelenkt und dann durch Mausklick gestartet werden. Die Messung startet automatisch beim ersten Nulldurchgang (vgl. Abbildung 4).

Das Simulationsexperiment unterscheidet sich daher hinsichtlich des notwendigen experimentellen Geschicks sowie hinsichtlich der Unsicherheiten der „Messungen“ vom dem echten Experiment. Beides sind jedoch typische Charakteristika eines Computerexperiments. Zum einen handelt es sich bei einer Computersimulation immer um eine modellhafte Abstraktion der Realität, die (meist) zwangsläufig zu einer einfacheren Bedienung führt. Zum anderen enthält eine Computersimulation eine analytische Lösung eines physikalischen bzw. naturwissenschaftlichen Problems, welche – bei gleichen Bedingungen – meist nur einen bestimmten „Messwert“ ausgibt.

4.2 FESTLEGUNG DER RELEVANTEN ZIELPOPULATION

Die aufgeführten Forschungsfragen sollen an einer Stichprobe von Schülern am Ende der Sekundarstufe I, d. h. der Klassenstufen 8 und 9, untersucht werden. Zum einen entspricht der so zu erwartende Altersdurchschnitt etwa den Stichproben der in Abschnitt 2.2 berichteten Studien zum Experimentieren und sichert so die Vergleichbarkeit der Ergebnisse. Zum anderen ist entsprechend den Bildungsstandards für das Fach Physik (KMK, 2004) festgelegt, dass Schüler am Ende Sekundarstufe I über eine Reihe von Kompetenzen aus dem Kompetenzbereich „Erkenntnisgewinnung“ verfügen, die für die erfolgreiche Durchführung des Experiments und die Auswertung der Messdaten notwendig sind. Beispiele solcher Kompetenzen sind etwa „Schülerinnen und Schüler werten gewonnene Daten aus“, „stellen an

einfachen Beispielen Hypothesen auf", „führen einfache Experimente [...] durch und werten sie aus" sowie „dokumentieren die Ergebnisse" (S. 11). Am Ende der Sekundarstufe I ist davon auszugehen, dass bei Schülern hinreichend Erfahrung bzgl. des Experimentierens im Physikunterricht vorhanden ist. Es kann ferner davon ausgegangen werden, dass ein Verständnis der wesentlichen physikalischen Größen (z. B. der Masse) sowie grundlegende mathematische Fähigkeiten, die für das Auswerten der Daten notwendig sind (z. B. Mittelwertbildung), vorhanden sind.

4.3 OPERATIONALISIERUNG DER KONSTRUKTE

Der folgende Abschnitt beschreibt, durch welche Verfahren und Instrumente die Konstrukte operationalisiert wurden. Zunächst werden dazu die methodologischen Vorüberlegungen und Entscheidungen im Hinblick auf die Operationalisierung der Verwendung der Argumentkategorien berichtet. Dabei wird ausgeführt, welche Randbedingungen zur Entwicklung eines eigenen Instruments zur Erfassung von Argumentationen beim Experimentieren geführt haben. Die Entwicklung dieses Instruments ist aus Gründen der Übersichtlichkeit in Anhang B dargestellt. Ferner wird in diesem Abschnitt die Operationalisierung der Hypothesen zum Fadenpendel sowie der persönlichen Faktoren durch die Verwendung etablierter Instrumente beschrieben. Eine übersichtliche Zusammenstellung der Itemtexte findet sich in Anhang C.

4.3.1 *Operationalisierung der Verwendung bestimmter Argumente beim Experimentieren*

Der überwiegende Teil naturwissenschaftsdidaktischer Forschung analysiert das Argumentieren auf der Basis von geschriebenen Texten (z. B. Kelly, Druker & Chen, 1998, 2007; Sandoval & Millwood, 2005) oder Interviews bzw. Videomitschnitten (z. B. Osborne, Erduran & Simon, 2004; Riemeier et al., 2012; Wächter & Kauertz, 2013). Dieser Ansatz ist gerechtfertigt, insbesondere dann, wenn die dialogische Komponente des Argumentierens untersucht werden soll. In der vorliegenden Arbeit werden jedoch Argumentationen in Bezug auf innersubjektive Entscheidungs- bzw. Lernprozesse, nämlich während des Verwerfens bzw. Aufstellens einer Hypothese beim selbstständigen Experimentieren untersucht. Diese Prozesse sind latenter Natur. Es ist daher ein observables Verhalten nötig, um die Verwendung bestimmter Argumente bei der Analyse von experimentellen Daten und Beobachtungen zu operationalisieren. Eine naheliegende Methode, nämlich die verbale oder schriftliche Beschreibung verwendeter Argumente durch den Lernenden nach dem Experimentieren, erscheint

in diesem Zusammenhang jedoch aus den folgenden Gründen problematisch:

Das in Abschnitt 2.3 entwickelte System zur Kategorisierung der von Lernenden gegebenden Argumente beim Experimentieren enthält Kategorien, von denen anzunehmen ist, dass sie bei einer direkten Befragung, z. B. während eines Interviews, Effekten sozialer Erwünschtheit unterliegen (Nederhof, 1985). Es ist vorstellbar, dass Schüler der 8. bzw. 9. Jahrgangsstufe bestimmte Argumente vermeiden, die aus ihrer Perspektive im Physikunterricht unangemessen erscheinen. Dies trifft z. B. auf die Argumentkategorien Intuition und Expertenwissen zu. Ein solcher Effekt würde zu Verzerrungen führen, die nur schwer zu kontrollieren sind.

Die Untersuchung der dargelegten Forschungsfragen setzt außerdem eine reliable Quantifizierung der Verwendung der Argumentationstypen voraus. Quantifizierung meint hier eine Messung auf mindestens Ordinalskalenniveau (vgl. Rost, 2003). Auch wenn prinzipiell aus Interviewtranskriptionen eine Quantifizierung abgeleitet werden kann, ist dieses Verfahren aus testökonomischen Gründen problematisch, da aufgrund der Komplexität der zu überprüfenden strukturellen Beziehungen ein relativ großer Stichprobenumfang nötig ist (vgl. Abschnitt 4.8.1). Bei Anwendung der Interviewmethode könnte eine Befragung nicht gleichzeitig in einer gesamten Klasse durchgeführt werden, sondern immer nur an einzelnen Schülern, so dass die Erhebung der Daten selbst zeit- und kostenintensiv wäre. Neben diesem ökonomischen Argument gibt es eine Reihe weiterer Einwände, die gegen die Quantifizierung qualitativer Daten sprechen (z. B. Hammer & Berland, 2014).

Im weiteren Verlauf der Arbeit wurde daher ein selbstauskunftsbasiertes Verfahren eingesetzt, dass nun, bei ansonsten identischem Ablauf der Untersuchung wie in Abschnitt 2.3, an die Stelle des Interviews rückt. Es ist bekannt, dass durch fragebogenbasierte Methoden Effekte sozialer Erwünschtheit weitgehend reduziert werden können (Nederhof, 1985; Richman, Kiesler, Weisband & Drasgow, 1999). Dieses Verfahren ist im Vergleich zu Interviews außerdem stärker standardisiert, da alle Probanden eine identische Itembatterie unter gleichen, kontrollierbaren Bedingungen beantworten. Das zugrundeliegende Itemkonzept dieser Arbeit beruht auf der Annahme, dass Probanden Aussagen vorgelegt werden, die auf die kognitive Berücksichtigung einer bestimmten Argumentkategorie während des Entscheidungsprozesses schließen lassen. Die Probanden werden dann aufgefordert anzugeben, inwiefern diese Aussage in ihrem eigenen Entscheidungsprozess eine Rolle gespielt hat. Unterschiede im manifesten Antwortverhalten lassen dann auf Unterschiede in der Verwendung bestimmter Typen von Argumenten bei der Entscheidung zum Wechseln bzw. Beibehalten der zuvor aufgestellten Hypothese schließen. Diese Art der Merkmalsdefinition ist in der Literatur als

„theoretisch begründete Definition eines Merkmals“ (Hartig, Frey & Jude, 2012, S. 145) bekannt. Aus der Art der Merkmalsdefinition ergeben sich Konsequenzen für den Beleg der Validität.

Als Antwortformat soll ein gebundenes Format in Form einer Ratingskala zum Einsatz kommen. Dies gibt den Probanden die Möglichkeit zu bewerten, in welchem Maße die vorgelegten Argumente bei ihrer eigenen Entscheidung berücksichtigt wurden.

Da zudem berücksichtigt werden muss, dass Probanden nach dem Experiment ihre eingangs aufgestellte Hypothese beibehalten, aber auch verwerfen können, müssen die Items immer so formuliert sein, dass sie in beiden Situationen beantwortbar bleiben, d. h., es müssen Aussagen vermieden werden, die sich explizit lediglich auf eine Situation beziehen. Die folgende Aussage ist ein Beispiel für einen Itemtext, der die o. g. Kriterien erfüllt.

Bei meiner Entscheidung zum Beibehalten/Verwerfen meiner Vermutung berücksichtige ich [...].

Anstelle des Platzhalters können nun Charakteristika der jeweiligen Argumentkategorie platziert werden.

Nachteil dieses Verfahrens ist, dass die Messung selbst, d. h. dass Lesen der Itemtexte, das Antwortverhalten möglicherweise beeinflusst. Die Vorteile (Vermeidung von Effekten sozialer Erwünschtheit, leichtere Quantifizierung und die Testökonomie) überwiegen diesen Nachteil jedoch.

Auf der Grundlage des dargelegten Konzepts wurde in einem dreischrittigen Vorgehen ein Test entwickelt, der die Stärke der Verwendung bestimmter Argumentkategorien erfasst: Zunächst wurde eine Itembatterie entwickelt, die im Folgenden in einer Expertenstudie in Bezug auf die inhaltliche Validität beurteilt wurde. Eine nachfolgende empirische Evaluation hatte zum Ziel, aus diesem inhaltlich validierten Set an Items nach psychometrischen Kriterien geeignete Items auszuwählen sowie empirische Evidenz für das Vorliegen von Konstrukt- bzw. faktorieller Validität vorzubringen. Weiterhin wurde geprüft, ob die Items die Voraussetzungen für sinnvolle Vergleiche von Mittelwerten aus verschiedenen Populationen erfüllen (hier der Vergleich der Gruppe der Probanden, die mit dem realen Experiment gearbeitet haben, mit der Gruppe der Probanden, die mit dem virtuellen Experiment gearbeitet haben). In der Literatur wird diese Prüfung der Übereinstimmung von Messmodellen in verschiedenen Populationen unter dem Begriff *Messinvarianz eines Tests* zusammengefasst (Brown, 2006; Kline, 2016; Vandenberg & Lance, 2000). Die Prüfung der Messinvarianz eines Tests ist im Kontext dieser Arbeit von besonderer Bedeutung, da die Probanden zunächst das Experiment durchführen müssen, um den Argumentationstest überhaupt sinnvoll beantworten zu können. Die zu operationalisierenden Konstrukte sind daher untrennbar mit der Experimentiersituation verknüpft. Folglich

muss eine wechselseitige Beeinflussung von Gruppenzugehörigkeit und Itemantworten während der Testentwicklung durch Prüfung der Messinvarianzbedingungen ausgeschlossen werden.

Die detaillierte Entwicklung dieses Testinstruments ist in Anhang B dargestellt. Es konnte ein Instrument entwickelt werden, das mit 20 Items die Verwendung der vier Argumentkategorien erfasst (fünf Items pro Subskala). Das Instrument zeichnet sich durch hohe inhaltliche, faktorielle, diskriminante und konvergente Validität sowie durch durch hohe Reliabilität aus. Ferner sind die überprüften Konstrukte invariant gegenüber der Testsituation, es liegt skalare Messinvarianz vor.

4.3.2 *Operationalisierung der Hypothesen*

Bei der Erfassung der Hypothesen wurde der zur Verfügung stehende Hypothesenraum bewusst eingeschränkt, um zu vermeiden, dass Probanden nicht prüfbare oder belanglose Hypothesen, z. B. ohne Verknüpfung zweier Variablen aufstellen (Hammann et al., 2006). In Form eines Multiple-Choice-Formats wurden daher lediglich die folgenden Hypothesen zur Auswahl vorgelegt (vgl. Kanari & Millar, 2004):

Vergrößern der Pendelmasse bewirkt, dass die Schwingungsdauer…

… größer wird.

… gleich bleibt.

… kleiner wird.

Unmittelbar nach dieser Abfrage wurden die Probanden mit einer offenen Frage aufgefordert, die Auswahl ihrer Hypothese zu begründen. Damit wurde versucht, die Probanden zu einer reflektierten Auswahl der Hypothese zu bewegen, da das Verfassen einer Begründung für die Wahl einer Hypothese zu einer intensiveren kognitiven Auseinandersetzung mit der Hypothese zwingt als das reine Ankreuzen. Diese Begründungen wurden nicht auf fachliche Richtigkeit hin analysiert und in der Auswertung nicht weiter berücksichtigt. Es ist anzunehmen, dass eine gewisse Ratewahrscheinlichkeit vorliegt. Es ist jedoch davon auszugehen, dass diese deutlich unter $1/3$ liegt, da die als korrekt zu bewertende Hypothese zum Zusammenhang zwischen Pendelmasse und Schwingungsdauer die Kovariation zwischen zwei physikalischen Größen negiert. Dies wird bei Schülern oft als nicht plausibel angenommen (Kanari & Millar, 2004).

Die von den Probanden aufgestellte Hypothese zum Thema ist die einzige Variable, die zu mehreren Zeitpunkten erfasst wurde. Entsprechend dem Untersuchungsdesign wurde die Hypothese zu drei Zeitpunkten erfasst: Unmittelbar vor dem Experiment, unmittelbar

nach dem Experiment sowie zur Follow-up-Untersuchung (vgl. Abschnitt 4.4).

4.3.3 *Operationalisierung der persönlichen Faktoren*

FACHWISSEN MECHANIK Zur Operationalisierung des Konstrukts „Fachwissen Mechanik" wurde auf ein Instrument auf der Basis von *single-select-multiple-choice*-Items von Zander (2016) zurück gegriffen (siehe auch Zander, Krabbe & Fischer, 2012). Zander (2016) orientierte sich bei Entwicklung dieses Instruments an aktuellen Kompetenzmodellen und den in den Bildungsstandards formulierten Anforderungsniveaus. Die inhaltliche Ausrichtung der Items erfolgte anhand der Lehrpläne der Länder sowie anhand von Schulbüchern. Inhaltlich deckt der Test die Themen *Kraft, Druck, mechanische Energie, Leistung* und *Innere Energie* ab. Zander (2016) unterzog die Items einer empirischen Evaluation und konnte so Evidenz für hinreichende Reliabilität sowie Validität erhalten; insbesondere konnte die Homogenität des Tests durch Etablierung eines Rasch-Modells gezeigt werden. Aus der 33 Items umfassenden Batterie von Zander (2016) wurden 22 Items ausgewählt, wovon 13 Items aus meist sprachlichen Gründen abgeändert wurden. Aufgrund dieser Änderungen wurde die Itemauswahl im Vorfeld der Untersuchung an einer Stichprobe von $n = 237$ pilotiert (Berndt, 2014). Es konnte gezeigt werden, dass die überarbeitete Itemauswahl den gängigen Testgütekriterien entspricht und daher weiterhin gut geeignet ist, das Fachwissen im Bereich Mechanik in der 8. und 9. Jahrgangsstufe zu erheben. Die Differenzierungsfähigkeit des Instruments ist trotz verkleinerter Itembatterie weiterhin hoch, da die Itemschwierigkeiten weiterhin die Personenschwierigkeiten gut abdecken. Die Reihung der Itemsschwierigkeiten entspricht den Ergebnissen von Zander et al. (2012).

KOGNITIONSBEDÜRFNIS Versuche zur Erfassung des Kognitionsbedürfnisses wurden erstmalig von Cacioppo und Petty (1982) unternommen. Aufgrund der Ergebnisse aus faktorenanalytischen Verfahren schlossen die Autoren auf eine eindimensionale Struktur des Konstrukts. Die Skala wurde in der Folge auf 18 Items reduziert (Cacioppo et al., 1984) und in verschiedenen Studien eingesetzt (für eine Übersicht siehe Cacioppo et al., 1996). Aufbauend auf der Arbeit von Cacioppo und Petty (1982) entwickelten Bless, Wänke, Bohner, Fellhauer und Schwarz (1994) eine aus 16 Items bestehende deutschsprachige Version der Skala, die in der vorliegenden Arbeit verwendet wurde. Allerdings zweifeln verschiedene Autoren die eindimensionale Struktur sowohl der englischsprachigen Version (Bors, Vigneau & Lalande, 2006; Furnham & Thorne, 2013; Hevey et al., 2012; Lord & Putrevu, 2006) als auch der deutsprachigen Version stark an (Pechtl, 2009). Insbesondere Hevey et al. (2012) konnten zeigen, dass die Ho-

mogenität der Skala stark durch einen Methodenfaktor der invers kodierten Items beeinflusst ist. Diesem Aspekt wird daher bei der Auswertung Rechnung getragen. Ferner wird kritisiert, dass die Itemtexte bisher äußerst unspezifisch und ohne konkreten Objekt- oder Situationsbezug formuliert sind, was allerdings an der ursprünglichen Definition des Konstrukts durch Petty und Cacioppo liegt (für eine detaillierte Diskussion der Schwächen siehe Pechtl, 2009). Beispielitems der hier verwendeten Skala nach Bless et al. (1994) sind etwa „Ich würde komplizierte Probleme einfachen Problemen vorziehen" bzw. „Denken entspricht nicht dem, was ich unter Spaß verstehe". Die Items wurden den Probanden zur Bewertung auf einer fünfstufigen Ratingskala mit den Ankern *stimme gar nicht zu / stimme wenig zu / stimme teils-teils zu / stimme ziemlich zu / stimme völlig zu* vorgelegt. Eine Übersicht der Itemtexte findet sich in Abschnitt C.1.2.

SITUATIONALES INTERESSE Das Konstrukt „Situationales Interesse" wurde mit einem an die Experimentiersituation angepassten Fragebogen nach Lewalter und Knogler (2014) bzw. Knogler, Harackiewicz, Gegenfurtner und Lewalter (2015) erhoben. Die Autoren orientierten sich bei der Entwicklung des Instruments an theoretischen Vorarbeiten zur Struktur des situationalen Interesses von Hidi und Renninger (2006), Krapp (2002) und Mitchell (1993). Dabei differenziert das Instrument zwischen der *catch*-Komponente, bestehend aus je einer aufmerksamkeitsorientierten Subskala (vier Items, z. B. „Die Experimentieraufgabe weckt meine Neugier."), einer emotionsorientierter Subskala (drei Items, z. B. „Die Experimentieraufgabe macht mir Spaß.") sowie der *hold*-Komponente, bestehend aus einer Subskala, die wertbezogene Elemente des situationalen Interesses erfasst (drei Items, z. B. „Die Beschäftigung mit der Experimentieraufgabe ist für mich nützlich.") und einer Subskala zur Erfassung epistemischer Komponenten (drei Items, z. B. „Über Teile der Aufgabe möchte ich gerne mehr erfahren."). Das situationale Interesse wird während der Bearbeitung des Fragebogens unmittelbar vor der Durchführung des Experiments erhoben, jedoch nachdem von den Versuchsleitern eine kurze Einführung in das Experiment gegeben wurde (bei der Simulation: Bedienung der Stoppuhr, Auslenkung des Pendels, Zurücksetzen auf Startbedingungen; bei dem Realexperiment Erklärung der Funktionen der Stoppuhr und des Aufbaus). Das so gemessene Konstrukt muss daher als das *erwartete* situationale Interesse bezeichnet werden, da die Probanden an dieser Stelle noch nicht das Experiment bearbeitet haben.

PERSÖNLICHE RELEVANZ Das Konstrukt „Persönliche Relevanz von Physik" wurde durch etablierte Skalen zur Erfassung der Werteinschätzung der Naturwissenschaften erhoben. Die Werteinschätzung der Naturwissenschaften ist abhängig vom jeweiligen Bezugsrahmen,

so dass etablierte Instrumente zwischen generellen (Bedeutung der Naturwissenschaften für Wirtschaft, Soziales und Gesellschaft), themenbezogenen (z. B. Umweltschutz, Atomkraft), persönlichen (persönliche Bedeutung für Alltag und Zukunft) sowie handlungsbezogenen Wertvorstellungen (Bedeutung der Naturwissenschaften auf persönliche Entscheidungsprozesse) unterscheiden (Knogler & Lewalter, 2014). Aufgrund der guten Passung zum ELM wurden die Skalen *Persönlicher Wert der Naturwissenschaften* („Ich finde, dass Physik mir hilft, die Dinge um mich herum zu verstehen.“) (Frey et al., 2009) und *Handlungsbezogenener Wert der Naturwissenschaften* („Physik hilft mir, vernünftige Entscheidungen zu treffen“) eingesetzt (Knogler und Lewalter, 2014; basierend auf Siegel und Ranney, 2003).

4.4 UNTERSUCHUNGSDESIGN

Bei den aus den Forschungsfragen 1 und 2 abgeleiteten Hypothesen handelt es sich um multiple Zusammenhangshypothesen, da sie einen Zusammenhang zwischen den vier persönlichen Faktoren und den vier Argumentkategorien sowie wiederum zwischen den vier Argumentkategorien und der Entscheidung zum Hypothesenwechsel und der Nachhaltigkeit dieser Entscheidung annehmen. Da diesen Hypothesen ein theoretisch fundiertes, gerichtetes Wirkmodell zugrundeliegt, werden im Folgenden die persönlichen Faktoren auch als Prädiktoren und die Argumentkategorien als Kriterien bezeichnet, deren Ausprägung durch die Prädiktoren erklärt werden soll (Döring & Bortz, 2016). Bei den aus Forschungsfrage 3 abgeleiteten Hypothesen handelt es sich um Unterschiedshypothesen, da sie Unterschiede zwischen den Ausprägungen der unabhängigen Variable Gruppe (real vs. virtuell) im Hinblick auf die vier abhängigen Variablen der Argumentkategorien postulieren. Alle angeführten Forschungsfragen haben explanativen Charakter, d. h. es sollen die aus der Theorie und den Vorarbeiten abgeleiteten Hypothesen auf ihre Gültigkeit hin geprüft werden.

Die notwendige quantitative Strategie wurde daher in einem experimentellen, einfaktoriellen Vergleichsgruppendesign mit Prä-, Post- und Follow-Up-Testung verfolgt, wobei sich die Messzeitpunkte Prä und Post auf die Durchführung des Experiments beziehen. Die Entscheidung zu diesem komplexen Design liegt in der Anlage der Forschungsfragen begründet:

Die Untersuchung der Forschungsfrage 3 (real vs. virtuell) macht ein experimentelles Vergleichsgruppen-Design unumgänglich, um die Unterschiede in den abhängigen Variablen kausal auf die Gruppenzugehörigkeit der Probanden zurückzuführen. Die persönlichen Faktoren können dabei als personenbezogene Störvariablen aufgefasst und als Kovariaten bei der Analyse berücksichtigt werden, wobei angemerkt werden muss, dass durch die vollständig randomisierte

Zuordnung der Probanden zu den Gruppen A-priori-Unterschiede bereits vermieden werden (Döring & Bortz, 2016). Im Rahmen dieses Designs können ebenfalls die multiplen Zusammenhangshypothesen aus Forschungsfrage 2 getestet werden. Um dem Einfluss der Argumentkategorien auf die Entscheidung zum Hypothesenwechsel und die Nachhaltigkeit dieser Entscheidung zu untersuchen, wurde dass Design auf der Variable Hypothese durch eine Messwiederholung ergänzt.

Anmerkung: Aus diesem komplexen Design ergibt sich eine weitere forschungsmethodische Implikation: Es ist denkbar, dass sich die in Forschungsfrage 2 untersuchten Einflüsse zwischen Prädiktoren und Kriterien in Richtung und Betrag zwischen den Gruppen unterscheiden. Dieser Sachverhalt muss bei der Auswertung geprüft werden.

4.5 ÜBERBLICK ÜBER DEN ABLAUF DER UNTERSUCHUNG

Im folgenden Abschnitt wird der Ablauf der Untersuchung überblicksartig zusammengefasst. Im Wesentlichen sind dabei drei Erhebungszeitpunkte zu unterscheiden; vor dem Experiment, nach dem Experiment sowie zwei bis drei Monate nach dem Experiment zur Follow-up-Erhebung (vgl. Abschnitt 4.8.2). Für die beiden Erhebungszeitpunkte unmittelbar vor bzw. nach dem Experiment wurde ein Fragebogen erstellt, der die Probanden durch die Untersuchung führt. Der Fragebogen ist in Abschnitt C.2 dokumentiert.

Zu Beginn der Untersuchung versammelten sich alle Schüler einer teilnehmenden Lerngruppe in einem der beiden zur Verfügung stehenden Räume. Die Probanden wurden über den Inhalt und die Fragestellungen der Untersuchung aufgeklärt. Es wurde zur Teilnahme motiviert und darauf hingewiesen, dass die Teilnahme an der Studie auf freiwilliger und pseudonymer Basis erfolgt. Den Probanden wurde daraufhin der Fragebogen ausgeteilt. Zusammen mit dem Untersuchungsleiter wurde die erste Seite des Fragebogens laut verlesen, aufkommende Fragen wurden gemeinsam geklärt. Die randomisierte Zuteilung der Lerngruppe zu den Experimentalgruppen erfolgte über einen Code auf der ersten Seite des Fragebogens, woraufhin eine Hälfte der Probanden zusammen mit einem weiteren Untersuchungsleiter den Raum wechselte. Danach bearbeiteten die Probanden eigenständig den Fragebogen und wurden so durch die Untersuchung geführt.

An dieser Stelle sei angemerkt, dass zu Beginn der Datenerhebung durch die Untersuchungsleiter beobachtet wurde, dass eine kleine Zahl von Probanden bereits beim Aufstellen der ersten Hypothese die Computersimulation bemühten. Diese Möglichkeit wurde im Laufe der Untersuchung unterbunden, indem die Computer mit einem Passwort gesperrt wurden (dieser Aspekt ist vermutlich ursächlich

für einen kleinen A-priori-Unterschied zwischen den Gruppen, vgl. Abschnitt 5.1).

Zur Charakterisierung der Stichprobe wurden zunächst einige Angaben zur Person erfragt. Daraufhin wurden die persönlichen Faktoren – mit Ausnahme des situationalen Interesse – erhoben. Die Probanden stellten nach der Einführung in die Aufgabenstellung eine Hypothese zum Zusammenhang von Masse des Pendelkörpers und Schwingungsdauer auf. Das situationale Interesses wurde erst nach Erläuterung der Aufgabenstellung, jedoch vor deren experimenteller Bearbeitung erhoben. So wurde sichergestellt, dass die Probanden zwar die experimentelle Aufgabenstellung kannten, das situationale Interesse aber nicht durch die Bearbeitung der Aufgabe in den beiden verschiedenen Experimentiergruppen (real vs. virtuell) beeinflusst wurde. Es ist nämlich denkbar und sehr plausibel, dass die Bearbeitung der Experimentieraufgabe zu Veränderungen im situationalen Interesse führt, die über die Gruppen hinweg unterschiedlich stark ausgeprägt sind. Dieser Zusammenhang war allerdings nicht Gegenstand der Untersuchung, außerdem konnte durch diese Entscheidung methodisch sichergestellt wurde, dass die Prädiktoren a priori keine Unterschiede aufweisen, da bei Gruppenvergleichen unter Berücksichtigung von Prädiktoren als Kovariaten die Prädiktoren zufällig verteilt sein sollten (Field, Miles & Field, 2012; Miller & Chapman, 2001).

Während der Einführung in die Aufgabenstellung wurde versucht, die Interessantheit durch die Einbettung der Thematik in einen Kontext zu steigern. Auf diese Weise sollten Bodeneffekte im Hinblick auf das situationale Interesse vermieden werden, da davon auszugehen ist, dass das physikalische Thema „Fadenpendel“ eine eher geringe Ausprägung des situationalen Interesses hervorruft.

Nachdem alle Probanden einer Teilgruppe an der jeweiligen Stelle des Fragebogens angekommen waren, konnte nach Aufforderung durch den Untersuchungsleiter mit dem Experiment begonnen werden. Die Probanden überprüften ihre eingangs aufgestellte Hypothese nun im Experiment. Nach eigenem Ermessen konnten die Probanden beliebig viele Messungen aufnehmen, bevor die Entscheidung über den Hypothesenwechsel erhoben wurde. Im unmittelbaren Anschluss an diese Entscheidung wurde die Verwendung der Argumentkategorien erhoben. Abschließend wurde den beteiligten Lehrkräften und den Probanden gedankt.

Die einzelnen Segmente des Fragebogens (vgl. Abschnitt C.2) sind hier noch einmal überblicksartig dargestellt:

1. Angaben zur Person
    a) Alter
    b) Klassenstufe
    c) Zeugnisnote Physik

   d) Gender
   e) Pseudonym
2. Persönliche Faktoren - Teil I
   a) Persönliche Relevanz
   b) Kognitionsbedürfnis
   c) Fachwissen
3. Experimentieraufgabe
   a) Einführung in die Aufgabe zum Experimentieren
   b) Hypothese vor dem Experiment
   c) Situationales Interesse
   d) Durchführung des Experiments
   e) Hypothese nach dem Experiment
4. Test zur Erfassung von Argumentationen beim Experimentieren

Um den Einfluss der Verwendung der Argumentkategorien auf die Nachhaltigkeit der Entscheidung bzgl. des Hypothesenwechsels zu erheben, wurde eine Follow-up Erhebung durchgeführt (vgl. Punkt 2). Da die Hauptuntersuchung innerhalb von vier Wochen vor den sechswöchigen Sommerferien erfolgte, fand die Follow-up-Erhebung nach dem Sommerferien statt. Zwischen der Durchführung des Experiments und der Nachfolgeerhebung liegen daher rund 60 bis 90 Tagen. Die Erhebung wurde von den Physiklehrerinnen und -lehrern der beteiligten Klassen während des Regelunterrichts durchgeführt. Der Follow-up-Fragebogen ist in Abschnitt C.3 dargestellt.

4.6 DATENAUFBEREITUNG

Die ausgefüllten Fragebögen wurden zunächst in tabellarische Form überführt und der relative Anteil der fehlenden Werte pro Item berechnet (für eine detaillierte Darstellung siehe Abschnitt C.4). Der durchschnittliche relative Anteil fehlender Werte liegt bei 1.38 %. Bei vier Items des Fachwissentests Mechanik zeigt sich, dass der Anteil fehlender Werte größer als 5 % ist (Items `K09P01`, `K06L01`, `K07P01`, `K13L01`). Insbesondere die Items `K09P01` und `K07P01` weisen einen relativen Anteil von $> 10\,\%$ an fehlenden Werten auf. Dies hat vermutlich inhaltliche Gründe, da beide Items das physikalische Konzept „vektorielle Addition von Kräften / Kraftpfeilen“ adressieren. Aufgrund dessen lässt sich vermuten, dass einem nicht kleinen Anteil der Probanden nötiges Wissen fehlt, um diese Aufgabe überhaupt zu bearbeiten und die beiden Aufgaben daher nicht „zufällig“ ausgelassen wurden. Dagegen spricht, dass das „Grundprinzip der Kräftezerlegung und -addition“ von Schülern der 7. und 8. Jahrgangsstufe

angewendet werden soll und das Themenfeld daher zum Untersuchungszeitpunkt bekannt sein sollte (Senatsverwaltung für Bildung, Jugend und Sport, 2006, S. 15). Aus diesem Grunde wurde der relativ hohe Anteil fehlender Werte bei diesen beiden Items ignoriert.

Aus den weiteren Analysen wurden 13 Datensätze ausgeschlossen, da diese Probanden nach Beginn der Untersuchung erschienen und in der Folge den letzten Teil des Fragebogens zur Erfassung der Argumentationen nicht bearbeiten konnten. Es wurden keine „Musterkreuzer" ausgeschlossen, da Muster zwar offensichtlich auftreten können, aber nur schwer kriteriengeleitet zu identifizieren sind.

Die Items des Fachwissentests wurden entsprechend der richtigen Lösung binär codiert. Alle fünfstufigen Ratingskalen wurden numerisch kodiert (1 = „trifft gar nicht zu" bis 5 = „trifft voll zu"). Die negativ formulierten Items der Skala Kognitionsbedürfnis wurden invertiert.

4.7 DATENANALYSE

Im Rahmen der vorliegenden Arbeit kamen verschiedene psychometrische und statistische Verfahren zum Einsatz, die im Folgenden kurz erläutert werden sollen. Besonderes Augenmerk gilt der Legitimierung methodischer Entscheidungen.

4.7.1 *Skalierung des Fachwissentests Mechanik mit dem einparametrischen Rasch-Modell*

Zur Auswertung des Fachwissentests Mechanik wurde ein probabilistisches Verfahren herangezogen. Das einparametrische Rasch-Modell (1PL) überführt dabei dichotome, binäre Itemantworten in Messwerte des dahinterstehenden latenten Merkmals anhand einer Wahrscheinlichkeitsfunktion, die über den Parameter der Itemschwierigkeit das Antwortverhalten und die Personenfähigkeit in Beziehung setzt. Die Modellparameter wurden durch das Marginal-Maximum-Likelihood (MML)-Verfahren geschätzt (Neumann, 2014; Rost, 2004).

Für die Schätzung der Personenfähigkeiten stehen verschiedene Verfahren zur Auswahl. „Die nach der WLE-Methode geschätzten Personenparameter sind die besten Punktschätzer der individuellen Messwerte" (Rost, 2004, S. 315). Daher wurden die Personenfähigkeiten mit dem Weighted-Likelihood-Estimator (WLE)-Verfahren nach Warm (1989) bestimmt.

4.7.2 *Analyse des komplexen Wirkungsmodells zum Argumentieren beim Experimentieren durch latente Strukturgleichungsmodelle*

Neben Standardverfahren kamen Verfahren der latenten Strukturgleichungsmodellierung zum Einsatz, welche technisch eine Erweiterung

der bereits während der Testentwicklung verwendeten konfirmatorischen Faktorenanalysen darstellen (vgl. Anhang B). Das Ziel von Strukturgleichungsmodellen (SGM) ist es, ein statistisches Argument dafür zu suchen, „dass ein vorgeschlagenes Theoriemodell auch empirisch sinnvoll ist“ (Urban & Mayerl, 2014, S. 14). SGM bieten gegenüber herkömmlichen Verfahren für die Analyse von komplexen Ursache-Wirksystemen verschiedene Vorteile, die hier kurz zusammengefasst sind (Hoyle, 2012; Kline, 2016; Urban & Mayerl, 2014):

- In SGM wird der Messfehler, d. h. die zufalls- oder methodenbedingte Varianz der Indikatoren auf *manifester* Ebene, explizit modelliert. Die *latenten* Variablen werden daher messfehlerfrei gemessen. „Dadurch können messfehler-bereinigte Schätzungen von freien Strukturparametern erreicht werden, wodurch [bei der] Analyse von Modellen mit latenten Konstrukten [...] die Reliabilität der Modellanalyse wesentlich erhöht wird“ (Urban & Mayerl, 2014, S. 16).

- Die Passung einer Modellschätzung zu den Daten kann über verschiedene absolute und deskriptive Maße beurteilt werden (vgl. Abschnitt E.1).

- Multiple Zusammenhangshypothesen (hier die Hypothesen aus Forschungsfragen 1 und 2) und Unterschiedshypothesen (die Hypothesen aus Forschungsfrage 3) können in einem Schritt in sog. Multigruppen-Strukturgleichungsmodellen (MG-SGM) getestet werden.

- Es besteht die Möglichkeit, verschiedene konkurrierende Modelle zu vergleichen (Green & Thompson, 2012).

- In SGM können auch nicht-multivariat-normalverteilte und nichtkontinuierliche Daten berücksichtigt werden.

- Verschiedene Voraussetzungen klassischer Verfahren (z. B. die Annahmer gleicher Varianzen und die Unkorreliertheit der Residuuen in ANOVAs) müssen in SGM nicht zwingend erfüllt sein bzw. können durch verschiedene Maßnahmen aufgehoben werden.

- Es gibt im Rahmen von SGM adäquate Möglichkeiten zum Umgang mit fehlenden Werten (s. u.).

Für eine ausführliche Zusammenfassung der Vorteile der Strukturgleichungsmodellierung siehe z. B., Urban und Mayerl (2014, S. 15).

In der Strukturgleichungsmodellierung werden das Strukturmodell, d. h. der Teil des Modells, der die latenten Faktoren untereinander durch direkte oder indirekte Effekte in Beziehung setzt (auch „Strukturteil“), von den Messmodellen eines oder mehrerer latenter

Faktoren unterschieden. Diese Messmodelle stellen die Beziehungen zwischen latenten Faktoren und manifesten Indikatoren (d. h. den Items) dar. Messmodelle und Strukturmodell bilden gemeinsam das Strukturgleichungsmodell (Kline, 2016; Urban & Mayerl, 2014). Es ist zudem zwischen endogenen und exogenen latenten Variablen zu unterscheiden. Exogene latente Variablen werden im Strukturmodell nicht durch andere latente Variablen erklärt (hier: Fachwissen Mechanik, persönliche Relevanz, situationales Interesse und Kognitionsbedürfnis), während endogene Variablen durch eben diese exogenen latenten Variablen erklärt werden sollen, d. h. hier die vier Argumentkategorien Evidenz, Messunsicherheiten, Intuition, und Expertenwissen (Kline, 2016).

Zur Analyse der Daten wurden in einem ersten Schritt die Messmodelle der exogenen und endogenen Variablen einzeln überprüft (Abschnitt 4.9). Zur Überprüfung des komplexen Wirkungsmodells Forschungsfragen 1 und 2 wurden die Messmodelle der vier exogenen Variablen bzw. der vier endogenen Variablen durch die aus dem ELM postulierten Wirkbeziehungen ergänzt. Zur Analyse der durch die Forschungsfrage 3 adressierten Unterschiedshypothesen kamen MG-SGM zum Einsatz. Während bei einfachen Strukturgleichungsmodellen die Mittelwerte der Indikatoren (engl. *intercepts*) normalerweise keinen Eingang in die Analyse finden (Urban & Mayerl, 2014), können MG-SGM die Mittelwerte berücksichtigen und gruppenspezifisch latente Mittelwerte mitschätzen und vergleichen (Brown, 2006; Kline, 2016; Urban & Mayerl, 2014).

AUSWAHL DES SCHÄTZVERFAHRENS Die Parameter, welche die Beziehungen zwischen latenten Faktoren und Indikatoren in den Messmodellen sowie zwischen den latenten Faktoren im Strukturmodell repräsentieren, müssen durch geeignete Verfahren bestimmt werden. Dabei ist zu prüfen, inwiefern die Voraussetzungen für diese Verfahren durch Charakteristika der Daten verletzt werden. Verletzungen der Voraussetzungen können in verzerrten Parameterschätzern und Standardfehlern resultieren (Finney & DiStefano, 2013). In der vorliegenden Arbeit werden fünfstufige Likert-Skalen zur Erfassung der personalen Faktoren verwendet (mit Ausnahme des Fachwissentests). Dabei handelt es sich streng genommen lediglich um Daten auf Ordinalskalenniveau. Zur Schätzung der Parameter bei ordinalen Daten wird in der Literatur in diesem Fall das Weighted-Least-Squares (WLS)-Verfahren empfohlen (Brown, 2006; Eid, Gollwitzer & Schmitt, 2013; Finney & DiStefano, 2013). Dieses Verfahren stellt jedoch hohe Ansprüche an den Stichprobenumfang, da in Form von Thresholds zwischen den Stufen der Likert-Skala zusätzliche Parameter bestimmt werden müssen. In der Entwicklung des Test zur Erfassung der Stärke der Verwendung der Argumentkategorien (vgl. Anhang B) lag jedoch nur ein vergleichsweise geringer Stichprobenum-

fang vor, der eine sichere Modellschätzung mit dem WLS-Verfahren nicht gewährleistet hätte.

Alternativ zu WLS-Verfahren können ML-Verfahren zur Parameterschätzung herangezogen werden. Das ML-Verfahren versucht die Modellparameter so zu schätzen, dass die beobachtete Stichprobenkovarianzmatrix eine maximale Wahrscheinlichkeit aufweist und bietet eine Reihe von nützlichen Eigenschaften: So können z. B. Standardfehler bzw. Konfidenzintervalle der Parameter bestimmt werden. Dadurch kann geprüft werden, ob sich Parameter signifikant von null bzw. voneinander unterscheiden (Eid et al., 2013). Eine Voraussetzung für ML-Verfahren ist das Vorliegen von multivariater Normalverteilung (Eid et al., 2013) bzw. kontinuierlichen manifesten Indikatoren. Likert-skalierte Items können per se nicht als kontinuierlich betrachtet werden und unterliegen daher nicht der Normalverteilung (Finney & DiStefano, 2013; Kline, 2016). Zudem konnte in der Testentwicklung bereits gezeigt werden, dass eine signifikante Verletzung der multivariaten Normalverteilung vorliegt. Aufgrund dieser Verletzungen der Skalen- und Verteilungsannahmen müssen sog. *robuste* ML- bzw. Bootstrapping-Verfahren zum Einsatz kommen (Brown, 2006; Bühner, 2010; Eid et al., 2013). Robuste ML-Verfahren liefern bei ordinalen Indikatoren mit fünf oder mehr Stufen auch bei Verletzung der Verteilungsvoraussetzungen asymptotisch unverzerrte Parameterschätzungen und Standardfehler (Beauducel & Herzberg, 2006; Finney & DiStefano, 2013; Pui-Wa & Qiong, 2012; Rhemtulla, Brosseau-Liard & Savalei, 2012). Im Gegensatz zu dem WLS-Verfahren können robuste ML-Verfahren außerdem fehlende Werte (unter Annahme der Missing-at-Random (MAR)-Bedingung) modellbasiert mit der Full-Information-Maximum-Likelihood (FIML)-Methode schätzen (z. B. Enders & Bandalos, 2001; Graham & Coffman, 2012; Urban & Mayerl, 2014). Aufgrund der dargelegten Vorteile des Verfahrens wurden alle Modelle in der vorliegenden Arbeit daher mit dem robusten ML-Verfahren berechnet. Dabei kam die FIML-Methode zum Einsatz, um adäquat mit fehlenden Werten umzugehen.

MODELLIDENTIFIKATION Alle im Rahmen dieser Arbeit geschätzten Modelle sind eindeutig bzw. überidentifiziert, da die Voraussetzung zur Modellidentifikation aufgrund der hohen Zahl an Indikatoren pro Faktor (> 3, keine oder nur sehr geringe Anzahl korrelierter Indikatorresidualvarianzen) stets erfüllt sind (Kline, 2016). Es wird daher im Einzelnen nicht weiter auf die Modellidentifikation eingegangen (zur Modellidentifikation siehe Kline, 2016, S. 203).

PRÜFUNG DER STATISTISCHEN SIGNIFIKANZ In der vorliegenden Arbeit wurden die unstandardisierten Parameterschätzer anhand von z-Werten auf das Vorliegen statistischer Signifikanz geprüft. Die Nullhypothese, dass der jeweilige Parameter null ist, muss auf dem

Tabelle 3: Übersicht über die verwendeten Cut-off-Kriterien zur Beurteilung eines *akzeptablen* Modellfits in CFA Faktorenanalysen und SGM. Eine tiefergehende Darstellung findet sich in Abschnitt E.1.

Gütemaß	Cut-off-Wert
CFI	$\geq .90$
RMSEA	$\leq .08$
SRMR	$\leq .08$

5%-Signifikanzniveau abgelehnt werden, wenn $|z| > 1.96$ (zweiseitig) (Kline, 2016, S. 51).

KONVERGENZ DER MODELLSCHÄTZUNG Alle in dieser Arbeit geschätzten Modelle konvergierten gegen eine zulässige Lösung, daher wird dieser Aspekt in der Darstellung der Ergebnisse nicht aufgegriffen.

FEHLENDE WERTE Analog zur Testentwicklungsstudie (vgl. Anhang B) wurde ein robustes FIML-Verfahren eingesetzt (z. B. Graham & Coffman, 2012), dass unter der MAR-Annahme auch bei unvollständigen Datensätzen unverzerrte Schätzer bestimmt (Lüdtke, Robitzsch, Trautwein & Köller, 2007; Rubin, 1976). Von der MAR-Annahme ist auszugehen, da der relative Anteil an fehlenden Werten in den Likertskalierten Items in der vorliegenden Arbeit äußerst gering ist (vgl. Abschnitt C.4).

BEURTEILUNG DES MODELLFITS Um die Passung des Modells zu den Daten zu beurteilen, können verschiedene absolute und relative Fitmaße herangezogen werden. In der vorliegenden Arbeit wurden die in Tabelle 3 dargestellten Cut-off-Kriterien zu Annahme eines Modells herangezogen. Eine Übersicht über die gängigsten Maße sowie eine ausführliche Darstellung findet sich in Abschnitt E.1.

BESTIMMUNG DER RELIABILITÄTEN DER SKALEN Bei der latenten Modellierung psychometrischer Tests kann die Reliabilität ρ als Verhältnis von wahrer Varianz $\text{Var}(T)$ zur Varianz der Messung $\text{Var}(Y)$ (Bühner, 2010; Eid et al., 2013) direkt bestimmt werden zu $\rho = \frac{\text{Var}(T)}{\text{Var}(Y)} = (\sum \lambda_i)^2 / ((\sum \lambda_i)^2 + \sum \Theta_{ii})$, wobei $\sum \lambda_i$ der Summe der unstandardisierten Faktorladungen und $\sum \Theta_{ii}$ der Summe der unstandardisierten Faktorvarianzen entspricht (Brown, 2006; Kline, 2016). Dieser Schätzer ist im Allgemeinen klassischen Alternativen wie Cronbachs α überlegen (Kline, 2016, S. 313). Hier und in der Testenwicklung Anhang B wurde daher dieses Verfahren zur Bestimmung der Reliabilität heran gezogen.

4.7.3 *Verwendete Software*

Die Analysen wurden mit der freien Statistiksoftware R (R Core Team, 2014) durchgeführt. Neben Standardpaketen wurden ergänzend die Pakete psych (Revelle, 2015) zur Berechnung gängiger psychometrischer Größen, `lavaan` (Rosseel, 5 2012) und `semTools` (Pornprasertmanit et al., 2014) zur Schätzung der Strukturgleichungsmodelle, ggplot2 (Wickham & Chang, 2015) zur Erzeugung von Diagrammen, reshape (Wickham, 2007) sowie plyr (Wickham, 2011) zum Aggregieren und Umformen der Datenstruktur, sowie TAM (Kiefer, Robitzsch & Wu, 2016) und WrightMap (Irribarra & Freund, 2014) für die Evaluation des Fachwissentests Mechanik mit dem Rasch-Modell genutzt.

Weil die Richtigkeit der aufgestellten Hypothesen nur binär kodiert werden kann, mussten logistische Regressionen zur Analyse der Daten im Hinblick auf Forschungsfrage 2 geschätzt werden. Da das Paket `lavaan`, das für die linearen Strukturgleichungsmodelle verwendet wurde, zurzeit nicht in der Lage ist, logistische Regressionen mit latenten Variablen zu schätzen, wurde dieser Teil der Fragestellung mit dem Softwarepaket `Mplus` (B. O. Muthén, 1998–2004) berechnet, um weiterhin auf die Vorteile latenter Modellierung zurückgreifen zu können. Zur Schätzung der Modelle mit nicht-linearem Teil in `Mplus` wurden identische Schätzverfahren wie in `lavaan` verwendet, zusätzlich wurde darauf geachtet, dass beim Übergang von `lavaan` nach `Mplus` die Messmodelle korrekt spezifiziert sind und beide Programme zu identischen Parametern führen.

4.8 STICHPROBE

4.8.1 *Stichprobenumfangsplanung*

Im Gegensatz zu klassischen statistischen Verfahren, wie der Varianzanalyse, bei denen manifeste Variablen Eingang in die Analyse finden, sind a priori durchgeführte Stichprobenumfangsplanungen bzw. Post-hoc-Poweranalysen im Kontext der hier eingesetzten latenten Strukturgleichungsmodellierung weitgehend unstandardisiert (Beaujean, 2014; Hancock, 2001) und Gegenstand aktueller Forschung (z. B. Wolf, Harrington, Clark & Miller, 2013). Es gibt Regeln zur Planung des Stichprobenumfangs, die z. B. aufgrund der Voraussetzungen für den Einsatz des ML-Verfahrens Stichproben von $100 < n < 200$ als ausreichend und Stichproben mit $n > 200$ als groß bezeichnen (Kline, 2011). Weitere Heuristiken setzen die Zahl der zu bestimmenden Parameter in ein Verhältnis zur Zahl der Fälle. Diese Praktiken werden aufgrund ihres unspezifischen Charakters weithin kritisiert, da meist die Komplexität und die Eigenschaften des Modells nicht in Betracht gezogen werden (Kline, 2016; Urban & Mayerl, 2014; Wolf et al., 2013).

Prinzipiell haben verschiedene Charakteristika der untersuchten Strukturgleichungsmodelle Einfluss auf die Teststärke bei einer gegebenen Stichprobengröße. Wolf et al. (2013) berichten, dass eine hohe Anzahl latenter Variablen mit einem größeren Stichprobenumfang einhergehen sollte. Zudem ist die Teststärke bzw. der Stichprobenumfang bei zu erreichender Teststärke auch von der Reliabilität der Instrumente (Urban & Mayerl, 2014) und insbesondere von der Art und Komplexität des zu untersuchenden Modells abhängig (Kline, 2016).

Bei der Bestimmung von Stichprobenumfang bzw. Teststärke sind prinzipiell zwei verschiedene Ansätze zu unterscheiden. Zum einen kann die Teststärke individuell für einzelne Parameter bestimmt werden. Dieses Verfahren ist jedoch bei großen Modellen relativ aufwendig (Hancock, 2001; Kline, 2016). Zum anderen existieren Vorschläge bei denen die Teststärke für die Evaluation auf Modellebene bestimmt wird (z. B. MacCallum, Browne & Cai, 2006, 1996). Die vorliegende Arbeit orientierte sich bei der Planung des Stichprobenumfangs anhand der zu erreichenden Teststärke auf Modellebene, da der Charakter der angeführten Forschungsfragen stärker das aus der Theorie abgeleitete Gesamtmodell adressiert als einzelne Parameter. Zudem ist aufgrund der Forschungslogik der Strukturgleichungsmodelle die Bestimmung der Teststärke auf Modellebene relevanter. Aus Perspektive des Forschers wird *für* die Nullhypothese „Modell entspricht den Daten“ getestet, welche hier gleichzeitig der Forschungshypothese entspricht. Falls also die Teststärke auf Modellebene nicht evaluiert wird, kann es sein, dass die Arbeit einem Bias unterliegt, da bei einem genügend geringen Stichprobenumfang Missspezifikationen des Modells erst gar nicht aufgedeckt werden können (MacCallum et al., 1996).

Das Verfahren nach MacCallum et al. (1996) bestimmt den Stichprobenumfang anhand vorgegebener Cut-off-Kriterien (für einen „test of not close-fit“ (S. 139) festgelegt zu $\epsilon_0 \leq .05$, $\epsilon_a \leq .01$) für den Root Mean Square Error of Approximation (RMSEA) eines SGM und anhand der Zahl der Freiheitsgrade (df) bei festgelegter Teststärke.[1]

Zum Zeitpunkt der Stichprobenumfangsplanung war die Auswahl der Instrumente noch nicht abgeschlossen, so dass in einer überschlägigen Rechnung die Zahl der Freiheitsgrade zu $\mathrm{df} > 500$ bestimmt wurde.[2] Nach MacCallum et al. (1996, S. 142) müssen bei $\mathrm{df} = 100$

1 Entsprechend gängiger Konventionen in der fachdidaktischen und psychologischen Forschung wurde der Stichprobenumfang mit dem Ziel einer zu erreichenden Teststärke von $1 - \beta = .80$ bei einem Signifikanzniveau von $\alpha = .05$ (Cohen, 1988) geplant.

2 Unter der großzügigen Annahme, dass acht latente Faktoren mit je fünf Variablen gemessen werden, ergeben sich $5 \cdot 8 = 40$ Indikatoren, so dass die Varianz-Kovarianzmatrix bereits aus $n = \frac{n \cdot (n+1)}{2} = 672$ Elementen besteht; dem gegenüber stehen etwa 120 zu bestimmende Parameter. Das Modell hat daher rund 550 Freiheitsgrade.

mindestens 90 Probanden (pro Gruppe) in die Untersuchung eingehen, um bei dem gegebenen Modell eine Teststärke von $1 - \beta = .80$ zuerreichen. Der somit ermittelte Stichprobenumfang liegt jedoch unter den genannten Mindestgrößen von idealerweise 200 Probanden pro Gruppe. Die hohe Teststärke bei dem zu testenden Modell auch bei kleinem Stichprobenumfang erklärt sich durch die hohe Zahl an manifesten Indikatoren (zur Diskussion siehe Kline, 2016; T. Lee, Cai, MacCallum & Hoyle, 2012; MacCallum et al., 2006). Um den Mindestvoraussetzungen Rechnung zu tragen, wurde daher mit 200 Probanden je Gruppe geplant und die Zahl um die 20% erhöht, was dem Teil der Probanden entspricht, die keine falsche Eingangshypothese aufstellen und von den nachfolgenden Analysen ausgeschlossen wurden (eine Begründung für diese Entscheidung findet sich in Abschnitt 5.2). Der Stichprobenumfang wurde daher auf $n \approx 500$ Probanden festgelegt. Für diese eher große Stichprobe spricht zudem, dass so ausreichend präzise Parameterschätzer erhalten werden können (Green & Thompson, 2012, S. 408). Aufgrund der dargelegten Argumente ist sichergestellt, dass eine ausreichend hohe Teststärke vorliegt und gleichzeitig den Mindestanforderungen der Strukturgleichungsmodellierung entsprochen wird.

4.8.2 *Stichprobenziehung und Charakteristika der Stichprobe*

Die Rekrutierung der Stichprobe erfolgte über bestehende Kontakte der Arbeitsgruppe und Partnerschulen der Humboldt-Universität zu Berlin. Es handelt sich daher um eine nicht-probabilistische Gelegenheitsstichprobe (Döring & Bortz, 2016). Die Schulleiter von vier Berliner Gymnasien erklärten sich bereit, an der Studie teilzunehmen. Es nahmen 38 Lerngruppen an der Untersuchung teil, wovon 18 Klassen die 8. Jahrgangsstufe und 20 Klassen die 9. Jahrgangsstufe besuchten. Insgesamt haben an der Untersuchung 938 Schülerinnen und Schüler teilgenommen[3]. Davon waren 53.4 % Mädchen, 0.6 % haben keine Angabe gemacht. Die Aufteilung über die Jahrgangsstufen ist homogen, 50.1 % der Probanden besuchen die 9. Jahrgangsstufe. Das Alter der Probanden liegt im Mittel bei 14.1 ($SD = 0.8$). Die Stichprobe setzt sich relativ gleichverteilt aus Schüler der vier Schulen zusammen (relativer Anteil an der Stichprobe je Schule liegt zwischen 19.7 % und 28.1 %). Die Schulnote „befriedigend“ ist der Median der letzten Physiknote.

An der Follow-up-Erhebung haben 816 Schüler teilgenommen. Davon konnten auf Grundlage des Pseudonyms 727 Schüler (77.5%)

3 Diese Zahl weicht deutlich vom geplanten Stichprobenumfang ab (vgl. Abschnitt 4.8.1). Dies ist der Tatsache geschuldet, dass in der Organisation ein Puffer einplant wurde, der letztlich nicht benötigt wurde. Hinzu kam, dass die Studie in den letzten Wochen vor Beginn der Sommerferien durchgeführt wurde. Da die durchführenden Personen zu dieser Zeit ständig in den Schulen anwesend waren, wurden oftmals Vertretungsstunden spontan mit der Untersuchung belegt.

zweifelsfrei den Datensätzen aus der ersten Erhebung zugeordnet werden.

Es ist plausibel anzunehmen, dass es sich hier um eine Klumpenstichprobe handelt, in der das Verhalten bzw. das Abschneiden der Probanden in den eingesetzten Testverfahren nicht unabhängig von der Zugehörigkeit zu einer Lerngruppe einerseits und ebenfalls nicht unabhängig von dem Besuch einer spezifischen Schule andererseits ist. Diese hierarchische Datenstruktur kann zu einer Unterschätzung der Standardfehler führen (z. B. Wild & Möller, 2014, S. 339) und ist in der empirischen Bildungsforschung häufig anzutreffen. In den weitern Analysen wird die hierarchische Struktur der Daten jedoch nicht modelliert. Diese Entscheidung kann u. a. darin begründet werden, da davon auszugehen ist, dass aufgrund der räumlichen Nähe der Schulen untereinander und soziogeografisch ähnlichen Gegebenheiten die Abhängigkeit auf Schulebene eher gering sein dürfte (zur Diskussion siehe auch Döring & Bortz, 2016).

Die dargelegten Charakteristika zeigen, dass die Stichprobe der festgelegten Zielpopulation entspricht (vgl. Abschnitt 4.2).

4.9 ANALYSE DER MESSMODELLE

Der folgende Abschnitt beschreibt die Evaluation der Messmodelle der persönlichen Faktoren und des Argumentationstests mittels konfirmatorischer Faktorenanalysen (CFA). Im Abschnitt C.5 findet sich dazu eine Zusammenfassung klassischer Skalenmaße sowie Histogramme der Likert-Skalen. Beides wird jedoch hier und im Ergebnisteil nicht weiter interpretiert. Die Bewertung der psychometrischen Qualität erfolgt einzig anhand der Evaluation der Messmodelle.

4.9.1 *Argumentkategorien*

Entsprechend der theoretischen Anlage des Tests zur Verwendung bestimmter Argumente beim Experimentieren (vgl. Anhang B) wurde die Passung eines vierfaktoriellen Modells evaluiert. Dieses Modell weist bereits einen akzeptablen Fit auf (Modell `arg`, Tabelle 4). Die Spezifikation von korrelierten Indikator-Residuen zwischen den vier jeweils aus der gleichen Subskala stammenden Indikatorpaaren `i1.18` und `i1.21`, `i1.03` und `i1.16`, `i6.08` und `i6.15`, `i7b.22` und `i7b.04` zeigt eine substanzielle Verbesserung des Modellfits (siehe Modell `arg2` in Tabelle 4). Diese Verbesserung des Modellfits kann mit einem χ^2-Differenzentest auf Signifikanz überprüft werden ($\Delta\chi^2(4) = 151.5$, $p < .01$).[4] Obwohl eine Reihe von Autoren ausdrücklich davor

4 Die Differenz des χ^2-Wertes der beiden Modelle in Tabelle 4 entspricht nicht der hier dargestellten Differenz. Dies ist darauf zurückzuführen, dass das hier verwendete Verfahren von Satorra und Bentler (2001) auf den unskalierten, d. h. nicht-robusten χ^2-Werten beruht (siehe auch Kline, 2016, S. 282).

Tabelle 4: Modellfit-Indizes für den Test zur Erfassung bestimmter Argumente beim Experimentieren

Modell	χ^2	df	p	χ^2/df	CFI	RMSEA [95 % C.I.]; p	SRMR
`arg`	619.7	164	$< .01$	3.78	.92	.05 [.05; .06]; .04	.06
`arg2`	428.1	160	$< .01$	2.68	.95	.04 [.04; .05]; 1	.06

Tabelle 5: Reliabilitäten für die Subskalen des Argumentationstests

Subskala	ρ
Intuition	.83
Expertenwissen	.75
Messunsicherheiten (explizit)	.73
Daten als Evidenz	.85

Hinweis: ρ ... CFA-modellbasierte Reliabilität

warnt, post-hoc Spezifikationen zur Modellverbesserung einzuführen (zur Diskussion siehe z. B. Brown, 2006, S. 140), ist dieses Vorgehen hier legitim, da es plausibel ist, dass ähnliche Wörter bei Items, die aus gleichen Subskalen entstammen, für eine gemeinsame Residualvarianz sind (Kline, 2016, S. 344).

Das resultierende Messmodell `arg2` wurde aufgrund des sehr guten Fits in den weiteren Analysen verwendet. Die modell-implizierten Reliabilitäten der Skalen sind in Tabelle 5 dargestellt. Darüber hinaus sind deskriptivstatistische Maße in Abschnitt C.5.1 dargestellt.

4.9.2 *Rasch-Analyse des Fachwissentests Mechanik*

Eine Auswertung der Daten auf manifester Ebene ergibt zunächst, dass durchschnittlich 59 % ($SD = 14\,\%, \min = 23\,\%, \max = 95\,\%$) der Items korrekt beantwortet werden. Die Antworten der Probanden wurden entsprechend binär kodiert (1 = richtige Antwort). Aus diesem Grund wurde hier zur weiteren Analyse wurde die Passung des einparametrischen Rasch-Modells evaluiert (für einen Überblick siehe z. B. Boone, Staver & Yale, 2014; Neumann, 2014). Dabei wurde der Mittelwert der Personenfähigkeitsverteilung auf null fixiert.

Die Itemschwierigkeiten liegen zwischen -3.13 und 2.88 Logits, die Standardabweichungen der Itemschwierigkeiten liegen zwischen 0.07 und 0.15. Die Mean-Square (MNSQ)-Fit-Statistiken liegen im Intervall von 0.91 bis 1.17 (Outfit) bzw. 0.94 bis 1.06 (Infit) und entsprechen damit verbreiteten Cut-off-Kriterien ($0.8 <$ Infit / Outfit < 1.2, (Bond & Fox, 2007; Neumann, 2014)). Die t-Statistiken der Itemparameter sind z. T. signifikant, können aber ignoriert werden, da die MNSQ-Fit-Statistiken innerhalb der Cut-off-Kriterien liegen (Boone et al., 2014, S. 166). Sämtliche Itemparameter und Fit-Statistiken sind in Tabelle 34 dargestellt. Eine grafische Darstellung der MNSQ-Fit-Statistiken finde sich in Abbildung 14.

Der Mittelwert der Itemschwierigkeiten liegt bei -0.44 Logits, die Varianz der latenten Personenfähigkeiten bei 0.36 Logits. In der Wright-Map sind die Personenfähigkeiten und Itemschwierigkeiten auf einer Skala aufgetragen (z. B. Boone et al., 2014) und in Abbildung 13 dargestellt. Die Wright-Map zeigt eine gleichmäßige Abdeckung der Skala im Bereich von etwa -2.0 Logits bis $+1.0$ Logits mit einer kleinen Lücke unterhalb von -1.4 Logits. Der durchschnittliche Abstand der Aufgaben beträgt in diesem Bereich 0.17 Logits. Im Rasch-Modell werden für die Bestimmung sowohl der Item- als auch der Personenfähigkeiten Reliabilitäten geschätzt (Bond & Fox, 2007). Die EAP-Reliabilität, welche die Zuverlässigkeit der Bestimmung der Itemparameter angibt wurde zu EAP $= .58$ bestimmt. Die WLE-Personenreliabilität, die am ehesten mit dem klassischen Schätzwert für die Reliabilität, Cronbachs α, verglichen werden kann, wurde zu $\rho_{\text{WLE}} = .56$ bestimmt. Der Mittelwert der WLE-Personenfähigkeit beträgt gemäß der Modellspezifikation null, die beobachtete Varianz der Personenfähigkeiten ist $\sigma^2_{\text{Personen}} = .63$ [5].

BEWERTUNG DES MESSMODELLS Es konnte das eindimensionales Rasch-Modell etabliert werden, das anhand der Fit-Statistiken der Items evaluiert wurde. Die Rasch-Konformität aller Items entspricht den Ergebnissen von Zander et al. (2012). Die Gültigkeit des Rasch-Modells ist gleichzeitig als ein Beleg für die Eindimensionalität bzw. Homogenität des zugrundeliegenden latenten Merkmals „Fachwissen Mechanik“ zu interpretieren (Boone et al., 2014). Die Differenz der Mittelwerte von Itemschwierigkeiten und Personenfähigkeiten lässt ferner den Schluss zu, dass der Test für die Stichprobe etwas zu leicht war. Dies zeigt sich auch an der Verteilung der Items in der Wright-Map. Daraus geht hervor, dass 14 Items eine Schwierigkeit von < 0 Logits aufweisen, d. h. dass 50 % der Probanden für diese Items eine Lösungswahrscheinlichkeit von mindestens 50 % aufweisen. Die Analyse der Wright-Map zeigt ferner, dass die Schwierigkeiten der Items recht gut über die Verteilung der Personenfähigkeiten verteilt sind.

5 Zur Unterscheidung von latenter Varianz und beobachteter Varianz siehe Rost (2004, S. 265)

Auffällig ist jedoch die Lücke im Bereich von 1.0 bis 2.9 Logits, der durch die Items `K06P03` und `K01P01` begrenzt wird. In diesem Bereich liegen 117 (12.4 %) der Probanden. Die große Lücke oberhalb von 1 Logit deutet darauf hin, dass der Test in diesem Bereich durch weitere Items ergänzt werden könnte, um auch in diesem Bereich das Fachwissen Mechanik differenziert aufzulösen. Diese Lücke hat vermutlich auch einen Einfluss auf die mit .58 relativ geringe Reliabilität des Tests. „Zu große Lücken könnten die Reliabilität des Messinstruments beeinträchtigen, weil sich eine größere Menge an Personen (mit möglicherweise unterschiedlichen Fähigkeiten) nicht unterscheiden lässt" (Neumann, 2014, S. 367). Auch die berichtete latente Varianz scheint eher gering zu sein, was jedoch vergleichbar mit anderen fachdidaktischen Instrumenten der Kompetenzmessung ist (Neumann (2014), siehe auch Kauertz (2008)). Im Hinblick auf die Reliabilität muss ferner diskutiert werden, dass mit den Aufgaben ein sehr heterogenes, Konstrukt gemessen wird, das sich über viele Subthemen erstreckt, die je nach Lehrkraft in verschiedenen Klassen unterschiedlich intensiv behandelt worden sein können bzw. für die Schüler unterschiedlich lange zurück liegen können. Die Heterogenität des Konstrukts in Kombination mit der relativ kurzen Testlänge von nur 22 Items kann ebenfalls für die relativ geringe Reliabilität des Tests verantwortlich sein (Bond & Fox, 2007).

INTEGRATION DES RASCH-PERSONENSCHÄTZERS IN DIE STRUKTURGLEICHUNGSMODELLE Im Gegensatz zu den Messmodellen der weiteren personalen Faktoren wurde der Mechaniktest mittels des Rasch-Modells evaluiert. Für die weiteren Analysen ist die individuelle Personenfähigkeit aus dem Rasch-Modell als *single*-Indikatorvariable in die weitere Strukturgleichungsmodellierung eingegangen. Konzeptuell wird dabei ein Pseudofaktordurch einen einzelnen Indikator gemessen. Da dieser Indikator (hier: der Personenschätzer aus dem Rasch-Modell) jedoch nicht ohne Messfehler gemessen wurde, wird der Teil der Indikatorvarianz, der nicht durch den latenten Faktor erklärt werden kann, d. h. die Residualvarianz des Indikators, im Modell auf einen a priori zu bestimmenden Wert beschränkt. Die Residualvarianz δ kann aus der beobachteten Varianz der Personenfähigkeit $\sigma^2_{\text{Personen}} = .63$ und der Reliabilität $\rho_{\text{WLE}} = .56$ bestimmt werden (vgl. Brown, Brown, S. 139 und Kline, Kline, S. 215).

$$\delta = \sigma^2_{\text{Personen}}(1 - \rho_{\text{WLE}}) = 0.28 \tag{21}$$

Das so resultierende Messmodell für den Mechaniktest besteht aus einem Pseudofaktor, welcher von einem Indikator mit festgelegter Residualvarianz gemessen wird und hat daher keine Freiheitsgrade ($\text{df} = 0$). Genau-identifizierte Modelle („just-identified model" Brown, 2006, S. 66) weisen einen perfekten Modellfit auf. Aus diesem Grund können hier keine Modellfit-Indizes angegeben werden (Urban & Mayerl, 2014, S. 77).

Tabelle 6: Fit-Indizes für die Messmodelle zur Skala Kognitionsbedürfnis. Das Modell `nfc-16Items` umfasst alle 16 Items der Skala nach Bless, Wänke, Bohner, Fellhauer und Schwarz (1994). Das Modell `nfc-6Items` umfasst nur die nicht-negativ formulierten Items.

Modell	χ^2	df	p	χ^2/df	CFI	RMSEA [95 % C.I.]; p	SRMR
`nfc-16Items`	574.9	104	$< .01$	5.53	.82	.07 [.06; .07]; $< .01$	.06
`nfc-6Items`	38.6	9	$< .01$	4.29	.97	.06 [.04; .08]; .17	.03

4.9.3 *Kognitionsbedürfnis*

Zur Evaluation der Skala Kognitionsbedürfnis nach Bless et al. (1994) (Itemtexte siehe Abschnitt C.1.2) wurden die 16 Items zunächst zu einem einfaktoriellen Messmodell spezifiziert. Das resultierende Messmodell weist zunächst einen nicht akzeptablen Fit auf (`nfc-16Items` in Tabelle 6). Wie schon in der Pilotierung der Skala (Berndt, 2014), deutet eine explorative Faktorenanalyse auf eine zweifaktorielle Lösung hin, bei der die negativ formulierten Items auf einen Methodenfaktor laden. Bors et al. (2006) sowie Hevey et al. (2012) berichten ebenfalls von einer Heterogenität der 16-Item Skala, die durch unterschiedlich formulierte Items verursacht wird. Hevey et al. (2012) favorisieren dabei ein zweifaktorielles Modell, das durch korrelierte Residualvarianzen der negativ formulierten Items dem Methodeneffekt Rechnung trägt, weisen jedoch auch darauf hin, dass die Forschungslage zur Interpretierbarkeit solcher Modelle unklar ist (für eine ausführliche Diskussion zum Thema siehe Meyer, 2010, S. 77). Dieser Argumentation folgend wurde in der vorliegenden Arbeit eine pragmatische Lösung zum Umgang mit dem Methodeneffekt gewählt, indem lediglich die sechs positiv formulierten Items zu einem Messmodell zusammengefasst wurden. Dieses Modell zeigt eine sehr gute Anpassungsgüte an die Daten (`nfc-6Items` in Tabelle 6) und ist daher in dieser Form in die weiteren Analysen eingegangen. Die modellimplizierte Reliabilität dieser verkürzten Skala beträgt $\rho = .78$.

Deskriptivstatistische Maße und das Histogramm der Skala, bestehend aus den sechs nicht-invertierten Items, sind in Abschnitt C.5.2 dargestellt.

4.9.4 *Situationales Interesse*

Weil den 14 Items der Skala Situationales Interesse (Itemtexte siehe Abschnitt C.1.3) inhaltlich vier Teilaspekte zugrundeliegen, wur-

Tabelle 7: Fit-Indizes für die Messmodelle des situationalen Interesses. Das einfaktorielle Modell fand Eingang in die weiteren Analysen.

Modell	χ^2	df	p	χ^2/df	CFI	RMSEA [95 % C.I.]; p	SRMR
`si-4fak`	297.5	59	$< .01$	5.04	.96	.07 [.06; .07]; $< .01$	.04
`si-1fak`	12.5	2	$< .01$	6.26	.99	.07 [.04; .11]; .09	.01

Tabelle 8: Latente Korrelationen zwischen den Komponenten des situationalen Interesses aus Modell `si-4fak`

	Aufmerksamkeit	Emotionale	Epistemische
Emotionale	.94		
Epistemische	.84	.70	
Wertbezogene	.84	.82	.66

de zunächst ein vierfaktorielles Modell evaluiert, das eine gute Passung zu den Daten zeigt (Modell `si-4fak` in Tabelle 7). Da in der vorliegenden Arbeit jedoch nur das Gesamtkonstrukt „situationales Interesse“ von Relevanz ist, wurde auf die Ebene der Subskalen hier verzichtet. Die Entscheidung kann ferner durch die starken Zusammenhänge zwischen den Subkonstrukten (Tabelle 8) legitimiert werden. Es wurden daher sogenannte *item-parcels* anhand des jeweiligen Durchschnitts der vier Subskalen gebildet und zu einem einfaktoriellen Messmodell zusammengefasst (zum Vergleich verschieden spezifizierter Messmodelle des situationalen Interesses siehe z. B. Knogler et al., 2015; zu den Vor- und Nachteilen des *item-parcellings* siehe Little, Rhemtulla, Gibson und Schoemann, 2013). Das resultierende einfaktorielle Modell weist einen guten Fit auf (Modell `si-1fak` in Tabelle 7). Die modellbasierte Reliabilität beträgt $\rho = .89$. Dieses Modell fand Eingang in die weiteren Anaylsen. Deskriptivstatistische Maße und Histogramme der Subskalen des Situationalen Interesses sind in Abschnitt C.5.3 zusammengestellt.

4.9.5 *Werteinschätzung der Naturwissenschaften*

Das Konstrukt Werteinschätzung der Naturwissenschaften wurde in der vorliegenden Arbeit durch die Skalen persönlicher und hand-

Tabelle 9: Fit-Indizes für das Messmodelle zum Konstrukt Werteinschätzung der Naturwissenschaften. Getestet wurde ein zweifaktorielles Messmodell (`wdn-2fak`), das die angenommene Substruktur repräsentiert, sowie ein einfaktorielles Modell (`wdn-1fak`), das alle verwendeten Indikatoren zusammenfasst.

	χ^2	df	p	χ^2/df	CFI	RMSEA [95 % C.I.]; p	SRMR
`wdn-2fak`	159.4	19	$< .01$	8.39	.91	.09 [.08; .10]; $< .01$	.04
`wdn-1fak`	65.3	19	$< .01$	3.44	.97	.05 [.04; .06]; .43	.03

lungsbezogener Wert der Naturwissenschaften operationalisiert (Abschnitt 4.3.3). Entsprechend wurde zunächst die Passung eines zweifaktoriellen Messmodells evaluiert (`wdn-2fak` in Tabelle 9). Dieses Modells verfehlt die Kriterien für einen noch akzeptablen Modellfit (RMSEA $> .08$). Auffällig ist die hohe Korrelation von $r = .90$ zwischen den beiden Subskalen. Dies kann als Indiz dafür gewertet werden, dass die beiden Subskalen „personaler“ und „handlungsbezogener“ Wert der Naturwissenschaften empirisch nicht zu trennen sind. Alle Items wurden daher in einem einfaktoriellen Messmodell zusammengefasst. Dabei wurde ein eine korrelierte Residualvarianz zwischen den Indikatoren `wdn.pers.2` und `wdn.pers.4` spezifiziert (Itemtexte siehe Abschnitt C.1.4). Dieses Vorgehen ist aus inhaltlichen Gründen legitimiert, da es sich bei diesen beiden Items um die einzigen Items der Skala handelt, die sich auf die Zukunft beziehen. Das resultierende Modell weist eine gute Passung auf (`wdn-1fak` in Tabelle 9). Die modellbasierte Reliabilität wurde zu $\rho = .77$ bestimmt.

4.9.6 *Analyse des globalen Messmodells*

In den vorherigen Abschnitten wurden die Messmodelle der Subskalen auf Skaleneben evaluiert. Kline (2016, S. 338) schlägt im sog. „two-step modeling“ vor, vor der Spezifikation regressiver Pfade zur Analyse der Forschungsfragen die einzelnen Messmodelle in ein globales Messmodells zu überführen und dessen Passung zu evaluieren. Nur bei einer akzeptablen Passung dieses Modells können dann strukturelle Regressionen eingeführt werden. In einem ersten Schritt wurden daher die Messmodelle der vier endogenen und vier exogenen latenten Variablen in einem globalen Messmodell aus acht latenten Faktoren zusammengefasst. Diese acht latenten Faktoren repräsentieren die theoretische Konzeption der gesamten Itembatterie (vier Skalen zur Erfassung der persönlichen Faktoren, davon das Fachwissen als

Tabelle 10: Fit-Indizes für die globalen Messmodelle. Neben dem theoretisch zu erwartendem achtfaktoriellen Messmodell (`glob-8fak`) wurde ein Modell geschätzt, bei dem alle Indikatoren durch einen einzigen Faktor erklärt werden (`glob-1fak`).

Modell	χ^2	df	p	χ^2/df	CFI	RMSEA [95 % C.I.]; p	SRMR
`glob-8fak`	1299.5	670	$< .01$	1.94	.94	.03 [.03; .04]; 1	.05
`glob-1fak`	6138.9	698	$< .01$	8.79	.49	.10 [.10; .10]; $< .01$	.12

single-item-Faktor, sowie vier Skalen mit denen die Stärke der Verwendung der Argumentkategorien erfasst wurde). Die Fit-Maße dieses Modells `glob-8fak` sind in Tabelle 10 dargestellt und liegen allesamt über den Cut-off-Kriterien. Da, abgesehen vom Fachwissentest, alle Items aus Likert-Skalen bestehen, wurde gegenüber diesem sehr komplexen Modell zudem ein einfaktorielles Messmodell geschätzt, um zu prüfen, ob dieses sparsamere Modell die empirischen Daten nicht besser abbildet (Kline, 2016). Die Fit-Maße dieses Modells deuten auf eine grobe Missspezifikation hin (`glob-1fak` in Tabelle 10). Ferner ist die Modellverschlechterung, bestimmt über einen χ^2-Differenzentest (Kline, 2016) signifikant ($\Delta\chi^2(28) = 3963.9$, $p < 0.01$). Weitere theoretisch denkbaren sinnvollen Messmodelle wurden nicht geschätzt. Damit ist die Passung des globalen Messmodells für die gesamte Stichprobe geprüft worden.

4.9.7 *Analyse der Messinvarianzbedingungen an der Stichprobe der Hauptuntersuchung*

Da sich die vorliegende Stichprobe in zwei Teilgruppen untergliedert (Probanden, die mit dem Realexperiment bzw. Probanden, die mit der Computersimulation gearbeitet haben), muss vor der Analyse der strukturellen Parameter die Invarianz der Messinstrumente geprüft werden. Zusammengefasst wird dabei wird geprüft, ob die Beziehungen zwischen den latenten Faktoren und den Indikatoren über die Gruppen hinweg identisch sind. Dabei wurde dasselbe Verfahren wie bereits zur Testentwicklung (Abschnitt B.7.2.2) verwendet. A priori ist jedoch davon auszugehen, dass es bei dieser Stichprobe zu keiner Verletzung der Invarianzbedingungen kommt: Zum einen wurden die persönlichen Faktoren vor dem Kontakt mit dem Experiment erhoben (die Probanden wurden zu diesem Zeitpunkt zwar schon randomisiert, haben das Experiment aber noch nicht durchgeführt), zum anderen wurde der für die vorliegende Arbeit entwickelte

Test zur Erfassung der Verwendung bestimmter Argumente während der Entwicklung auf das Vorliegen von Messinvarianz geprüft (vgl. Abschnitt 4.3.1).

Analog zu Abschnitt B.7.2.2 wurden zunächst die Messmodelle separat in jeder Gruppe geschätzt und dann schrittweise das Vorliegen von konfiguraler, metrischer und skalarer Invarianz geprüft. Als Beurteilungskriterien für den Vergleich der hierarchisch geschachtelten Modelle gelten die in Abschnitt E.1.2 dargelegten Kriterien. Die Ergebnisse der Modellschätzungen und -vergleiche sind in Tabelle 11 zusammengefasst. Die Fit-Indizes für die Messmodelle, basierend auf den beiden Subsamples, liegen über allen Cut-off-Kriterien. Im Folgenden wurden daher die Multigruppen-Modelle geschätzt, in denen sukzessive einzelne Parameter restringiert, d. h. für beide Gruppen auf einen identischen Wert gesetzt werden. Auch diese Modelle weisen einen guten Modellfit auf. Beim Übergang vom konfigural invarianten zum metrisch invarianten Messmodell zeigt sich keine signifikante Verschlechterung (siehe vorletzte Zeile in Tabelle 11), auch die Differenzen der approximativen Fit-Indizes liegen unterhalb der Cut-off-Werte. Beim Übergang vom metrisch-invarianten zum skalarinvarianten Messmodell zeigt der $\Delta\chi^2$-Differenzentest eine signifikante Verschlechterung des Fits auf, ΔCFI und ΔRMSEA sind jedoch äußerst gering. Daher kann davon ausgegangen werden kann, dass der χ^2-Differenzentest aufgrund der großen Stichprobe und damit verbundenen hohen Teststärke signifikant wird, der Effekt der Abweichung aber marginal ist (zur Beurteilung hierarchisch geschachtelter Modelle siehe Abschnitt E.1.2).

Tabelle 11: Analyse der Messinvarianzbedingungen für das globale Messmodell, bestehend aus acht latenten Faktoren

Modell	χ^2	df	p	χ^2/df	CFI	RMSEA [95 % C.I.]; p	SRMR	$\Delta\chi^2$	Δdf	p	ΔCFI	ΔRMSEA
`Gruppe R`	1037.8	670	$< .01$	1.55	.930	.040 [.030; .040]; 1	.05					
`Gruppe V`	104.6	670	$< .01$	1.55	.930	.040 [.030; .040]; 1	.06					
`konfigural`	2078.4	1340	$< .01$	1.55	.931	.037 [.034; .040]; 1	.06					
`metrisch`	2114.5	1371	$< .01$	1.54	.930	.036 [.033; .039]; 1	.06					
`skalar`	216.8	1402	$< .01$	1.54	.929	.036 [.033; .039]; 1	.06					
Modellvergleiche												
`konfigural` vs. `metrisch`								36.50	31	.23	$-.001$	$-.001$
`metrisch` vs. `skalar`								46.30	31	.04	$-.001$	.000

Hinweise: `Gruppe R` ... Messmodell separat in Gruppe Realexperiment; `Gruppe V` ... Messmodell separat in Gruppe Computersimulation; `konfigural` ... MG-CFA-Modell mit gleicher Faktorladungsstruktur (konfigurale Messinvarianz); `metrisch` ... MG-CFA-Modell mit identischen Faktorladungen in beiden Gruppen (metrische Messinvarianz); `skalar` ... MG-CFA-Modell mit gleichen Faktorladungen und gleichen Itemintercepts (skalare Messinvarianz)

4.9.8 *Zwischenfazit Messmodelle*

Bei den Instrumenten zur Erfassung der personalen Faktoren handelt es sich um elaborierte Instrumente, deren Reliabilität und faktorielle Validität im Rahmen der konfirmatorischen Faktorenanalysen (CFA) an dieser Stichprobe erneut geprüft wurden. Die CFAs deuten dabei auf eine überwiegend gute Passung zwischen Daten und angenommener Modellstruktur hin. An den Messmodellen der personalen Faktoren Kognitionsbedürfnis und Werteinschätzung der Naturwissenschaften mussten post hoc Modifikationen vorgenommen werden. Es zeigte sich ein Methodeneffekt aufgrund invertierter Items im Messmodell für das Kognitionsbedürfnis. Im Messmodell Werteinschätzung der Naturwissenschaften konnte die Einführung korrelierter Indikatorresiduen in das Messmodell eine signifikante Modellverbesserung leisten. Der Fachwissentest weist eine relativ geringe Reliabilität auf. Mögliche Gründe dafür wurden dargelegt. Der Test, mit dem die Stärke der Verwendung der Argumentkategorien erfasst wurde, weist auch an dieser Stichprobe eine hohe Reliabilität, faktorielle Validität sowie Messinvarianz auf (vgl. Anhang B). Auch das globale Messmodell, dass alle 16 Subskalen enthält, zeigt eine gute Passung zu den Daten. Es ist davon auszugehen, dass die verwendeten Instrumente für die Beantwortung der Forschungsfragen eine hinreichend hohe Testgüte aufweisen.

5 ERGEBNISSE

Nach einem Überblick über die von den Probanden aufgestellten Hypothesen zum Fadenpendel wird das Datenmaterial analyisiert. Es werden die Hypothesen zum Einfluss der personalen Faktoren auf die Stärke der Verwendung der Argumentkategorien geprüft (Forschungsfrage 1) und die Unterschiede beim Argumentieren zwischen den beiden Gruppen – real vs. virtuell – bestimmt (Forschungsfrage 3). Abschließend wird der Einflusses der Stärke der Verwendung der Argumentkategorien auf die fachliche Richtigkeit der nach dem Experiment aufgestellten Hypothesen analysiert (Forschungsfrage 2).

Der Aufbau aller Teilabschnitte in diesem Kapitel ist identisch. Es werden zunächst die Parameter der verschiedenen statistischen Modelle und Tests berichtet. Da zur Auswertung mitunter unterschiedliche Verfahren verwendet werden können, werden teilweise die Ergebnisse von mehr als einem Verfahren im Fließtext (oder im Anhang) berichtet. Obwohl dieser Ansatz leichte Redundanz erzeugt, wird so die Tragfähigkeit der Ergebnisse erhöht und die Stabilität der Parameterschätzungen verdeutlicht. Daran schließt sich jeweils die Bewertung der statistischen Hypothesen an. Bei allen in Abschnitt 3.2 formulierten Hypothesen handelt es sich um theoretisch begründbare gerichtete Alternativhypothesen. Alle berichteten statistischen Hypothesentests testen jedoch immer die Nullhypothese, nicht die Alternativhypothese. Wird ein signifikanter Effekt oder Unterschied berichtet, bezieht sich dies immer auf die Nullhypothese, die den entsprechenden Effekt negiert, nicht auf die Alternativhypothese. Bei der Bewertung der Hypothesen wird jedoch im Folgenden aus Gründen der Lesbarkeit die Bewertung der Nullhypothese ausgelassen und stattdessen die Alternativhypothese bewertet. Im letzten Abschnitt dieses Kapitels findet sich eine Zusammenfassung der getesteten Hypothesen (Abschnitt 5.6).

5.1 IM VERLAUF DER UNTERSUCHUNG AUFGESTELLTE HYPOTHESEN ZUM FADENPENDEL

Im Laufe der Untersuchungen stellten die Probanden zu drei Zeitpunkten eine Hypothese zum Thema auf. Abbildung 6 stellt in Form eines Flussdiagramms diese Entscheidungen dar. Exemplarisch sei hier ein Pfad näher beschrieben: Zunächst stellen 88 % der Probanden vor dem Experiment eine fachlich inkorrekte Hypothese auf (vgl. Anhang B), d. h. es wird ein positiver oder negativer Zusammenhang zwischen Pendelmasse und Schwingungsdauer angenommen.

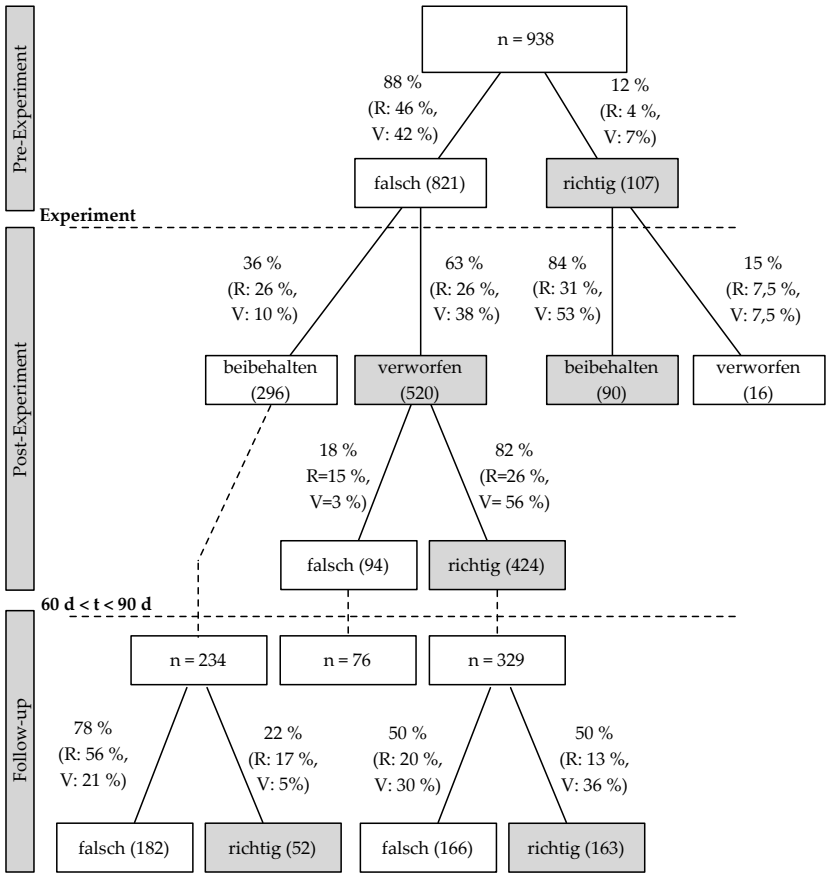


Abbildung 6: Aufgestellte Hypothesen im Laufe der Untersuchung. Dargestellt sind die drei Messzeitpunkte Prä- und Post-Experiment sowie Follow-up. Relative Angaben beziehen sich auf den vorhergehenden Knotenpunkt (R . . . Realexperiment; V . . . Simulationsexperiment). Abweichungen zwischen einzelnen Ebenen sind fehlenden Werten geschuldet.

63 % der Probanden, die eine fachlich inkorrekte Hypothese aufstellen verwerfen diese nach dem Experiment zugunsten einer anderen Hypothese, wobei 82 % dann die fachlich korrekte Hypothese aufstellen. Von 821 Probanden mit falscher Eingangshypothese stellen daher rund 424, d. h. rund 52 % nach dem Experimentieren eine korrekte Hypothese auf. Von dieser Teilstichprobe konnten in der Follow-up-Erhebung 329 Probanden erreicht werden, wovon jedoch nur rund jeder zweite Proband erneut die fachlich adäquate Hypothese aufstellt.

Augenscheinlich ist zu erkennen, dass die relativen Häufigkeiten der Entscheidungen über die Gruppen hinweg ungleich erscheinen. Mit zweidimensionalen χ^2-Tests nach Pearson (Kontingenzanalyse) wurde geprüft, ob die Wahl der Hypothese zu den verschiedenen Zeitpunkten unabhängig von der Experimentierumgebung ist. Diese Nullhypothese muss für die Wahl der Hypothese vor dem Expe-

riment verworfen werden ($\chi^2(1) = 6.56, p = .01$). Die Effektstärke dieses Zusammenhangs ist mit $\phi = .08$ jedoch als äußerst gering einzustufen (Cramers ϕ ist ein Effektstärkemaß für $2x2$-Analysen; zur Interpretation siehe Abschnitt E.2). Dieser A-priori-Unterschied ist schwer zu erklären, da die Probanden randomisiert den Gruppen zugeordnet wurden und zu dem Zeitpunkt das Experiment noch nicht durchgeführt haben. Eine mögliche Erklärung könnte sein, dass Probanden bereits beim Aufstellen der ersten Hypothese kurz mit der Computersimulation experimentiert haben (vgl. Abschnitt 4.5). Aufgrund der geringen Größe dieses Effekts wird dieser A-priori-Unterschied in den weiteren Analysen vernachlässigt. Weiterhin zeigt sich sowohl für die Entscheidung zum Beibehalten bzw. Verwerfen der falschen Hypothese ($\chi^2(1) = 69.8, p < .001, \phi = .27$), als auch für die Richtigkeit der nach dem Experiment aufgestellten Hypothese ($\chi^2(1) = 81.0, p < .001, \phi = .29$), ein mittelstarker signifikanter Zusammenhang zwischen Lernumgebung und Ausgang der Hypothesenwahl. Dieser Effekt zeigt sich auch für die Richtigkeit der Hypothese zum Zeitpunkt der Follow-up-Untersuchung ($\chi^2(1) = 6.0, p = .01, \phi = .08$), jedoch ist hier nur von einem kleinen Effekt zu sprechen.

Diese Ergebnisse werden jedoch hier und im Folgenden nicht weiter diskutiert, da die Forschungsfragen (Abschnitt 3.2) diesen Sachverhalt nicht explizit adressieren, sondern die Argumentation als mediierenden Prozess zwischen dem Typ der Lernumgebung und dem Lernerfolg in Form des Aufstellens einer richtigen Hypothese annehmen.

5.2 FESTLEGEN DER TEILSTICHPROBEN FÜR DIE WEITEREN ANALYSEN

Wie in Abbildung 6 dargestellt, ergeben sich im Laufe der Untersuchung verschiedene Subsamples, die durch unterschiedliche Entscheidungen bzgl. der Wahl der Hypothese festgelegt werden können. Von Bedeutung ist die Teilgruppe der Probanden, die vor dem Experiment eine falsche Hypothese aufgestellt haben. In allen weiteren Analysen findet lediglich die Teilstichprobe der 821 Probanden Eingang, die vor dem Experiment eine falsche Hypothese aufgestellt haben. Von den Probanden, die im Zuge der Follow-up-Erhebung erreicht und zugeordnet werden konnten, haben 639 Probanden zu Beginn eine falsche Hypothese aufgestellt und fanden daher Eingang in die weiteren Analysen. Der Teil von Forschungsfrage 2, der die Nachhaltigkeit des Lernerfolgs adressiert, wird anhand dieser Stichprobe analysiert. Die Entscheidung zum Ausschluss der Probanden mit fachlich korrekter Eingangshypothese wird inhaltlich legitimiert durch die Tatsache, dass die Verarbeitung der experimentellen Daten und Beobachtungen bei Probanden, die eine fachlich adäquate Hypothese aufgestellt haben und dementsprechend keinen kogniti-

ven Konflikt erfahren, qualitativ einen anderen Charakter aufweist, da die Informationen kongruent zur Vermutung sind. Dieses Vorgehen entspricht dem in Abschnitt 2.1.5 dargelegten Ansatz, nach dem die Konfrontation mit nicht-hypothesenkonformen Daten ein Auslöser für eine intensive Argumentation sein kann.

5.3 ANALYSE DES EINFLUSSES VON PERSONALEN FAKTOREN AUF DIE VERWENDUNG DER ARGUMENTE

Der folgende Abschnitt analysiert den Einfluss der persönlichen Faktoren Fachwissen, Kognitionsbedürfnis, situationales Interesse und Werteinschätzung der Naturwissenschaften auf die Verwendung der Argumentkategorien Intuition, Expertenwissen, Messunsicherheiten zentral und Evidenz (Forschungsfrage 1). Zur Analyse dieses komplexen Wirkzusammenhangs wurde im Rahmen der Strukturgleichungsmodellierung ein multiples Regressionsmodell spezifiziert. Die Entwicklung dieses Modells wird im folgenden Abschnitt dargelegt.

Für die Entwicklung des Regressionsmodells sind zum einen die Korrelationen zwischen den persönlichen Faktoren und den Argumentkategorien als erste Indizien von Interesse. In dem globalen Messmodell aller acht verwendeten Skalen (`glob-8fak` in Tabelle 10) werden keine strukturellen Vorgaben zwischen den latenten Variablen gemacht. Daher können hier die modell-implizierten bivariaten Korrelationen zwischen den latenten Variablen bestimmt werden. Diese Korrelationen sind in Tabelle 12 dargestellt.

Bedeutsam sind dabei zum einen die kleinen bis mittleren negativen Korrelationen zwischen der Argumentkategorie Intuition und den persönlichen Faktoren sowie kleine positive Zusammenhänge zwischen der Argumentkategorie Daten als Evidenz und den persönlichen Faktoren. Es zeigt sich zudem ein mittlerer negativer Zusammenhang zwischen der Verwendung der Kategorie Daten als Evidenz und Intuition sowie ein mittlerer positiver Zusammenhang zwischen der Kategorie Daten als Evidenz und Expertenwissen (zur Interpretation von Effektstärken siehe Abschnitt E.2).

Zum anderen sind die Zusammenhänge zwischen den exogenen Variablen von Bedeutung (unterer rechter Quadrant in Tabelle 12), da diese als Hinweis für das Vorliegen von Multikollinearität, d. h. „die gegenseitige lineare Abhängigkeit von zwei oder mehr exogenen Variablen“ (Urban & Mayerl, 2014, S. 44), gewertet werden können. Multikollinearität kann zu verzerrten Parameter-Schätzern und überhöhten Standardfehlern führen. Urban und Mayerl (2014) nennen als Cut-off-Kriterium eine Korrelation von $|r| > .80$. Es zeigen sich starke lineare Zusammenhänge zwischen den Variablen Kognitionsbedürfnis und Fachwissen ($r_{\text{fw}\sim\text{nfc}} = .51$), Kognitionsbedürfnis und situationales Interesse ($r_{\text{nfc}\sim\text{si}} = .58$), Kognitionsbedürfnis und der Werteinschätzung der Naturwissenschaften ($r_{\text{nfc}\sim\text{wdn}} = .74$), so-

Tabelle 12: Zusammenhänge zwischen den latenten Variablen. Dargestellt sind modell-implizierte bivariate Korrelationen. Zur besseren Orientierung sind Hilfslinien eingezeichnet, a) die Korrelationen zwischen den Argumentkategorien (oberer linker Quadrant), b) Zusammenhänge zwischen den Prädiktoren und den Argumentkategorien (unterer linker Quadrant), Zusammenhänge zwischen den Prädiktoren (unterer rechter Quadrant).

	int	exp	mu	evi	fw	nfc	si
exp	.00						
mu	.15*	.08					
evi	−.44*	.42*	.05				
fw	−.37*	.10	.11	.25*			
nfc	−.25*	.10	.21*	.20*	.51*		
si	−.14*	.13*	.20*	.16*	.23*	.58*	
wdn	−.15*	.18*	.30*	.19*	.41*	.74*	.65*

* $p < .05$; int ... Argumentkategorie Intuition; exp ... Argumentkategorie Expertenwissen; mu ... Argumentkategorie Messunsicherheiten (explizit); evi ... Argumentkategorie Daten als Evidenz; fw ... Fachwissen Mechanik; nfc ... Kognitionsbedürfnis; wdn ... Werteinschätzung der Naturwissenschaften; si ... Situationales Interesse.

wie zwischen dem situationalen Interesse und der Werteinschätzung der Naturwissenschaften ($r_{\text{si}\sim\text{wdn}} = .65$). Ferner zeigt sich ein mittlerer Zusammenhang zwischen Fachwissen und der Werteinschätzung der Naturwissenschaften sowie ein kleiner Zusammenhang zwischen dem Fachwissen und dem situationalen Interesse. Keiner der Werte übersteigt das von Urban und Mayerl (2014) vorgeschlagene Kriterium. Es wird daher in den folgenden Analysen davon ausgegangen, dass die Multikollinearität zwischen den Prädiktoren vernachlässigt werden kann.

Im nachfolgenden Schritt wurden entsprechend der Forschungsfrage 1 die Verwendung der Argumentkategorien auf die persönlichen Faktoren regressiert. Dazu wurden im Strukturmodell die entsprechenden Pfade so spezifiziert, dass die Argumentkategorien als abhängige Variablen durch die persönlichen Faktoren erklärt werden. Technisch handelt es sich bei diesem Vorgehen um eine multiple Regression der Argumentkategorien auf die persönlichen Faktoren als Prädiktoren (Cohen, Cohen, West & Aiken, 2003). Zur Veranschaulichung ist dieses hypothetisierte Strukturmodell in Abbildung 7 grafisch dargestellt.

Anstelle der Korrelationen zwischen den vier exogenen und den vier endogenen Variablen werden nun partielle Regressionskoeffizienten zur Vorhersage der Verwendung der vier Argumentkategorien durch die vier persönlichen Faktoren geschätzt. Diese direkten Ef-

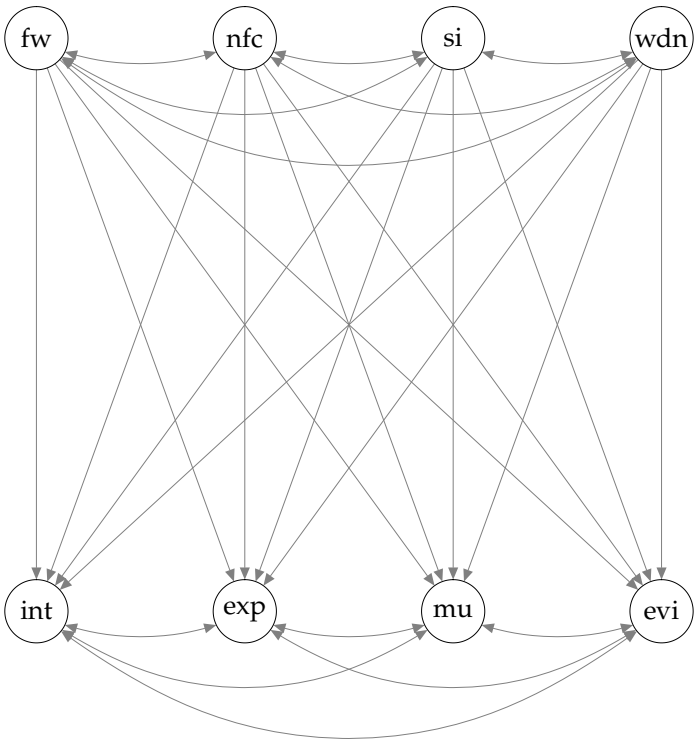


Abbildung 7: Hypothetisiertes Strukturmodell zum Wirkungszusammenhang zwischen den persönlichen Faktoren und der Verwendung bestimmter Argumente. Die persönlichen Faktoren als exogene, d. h. nicht durch das Modell erklärte Variablen sind oben dargestellt (fw ... Fachwissen, nfc ... Kognitionsbedürfnis, si ... Situationales Interesse, wdn ... Werteinschätzung der Naturwissenschaften). Die Argumentkategorien als endogene, d. h. durch das Modell erklärte Variablen, sind im Diagramm unten dargestellt (int ... Intuition, exp ... Expertenwissen, mu... Messunsicherheiten (explizit), evi ... Daten als Evidenz). Doppelpfeile repräsentieren zugelassene Korrelationen zwischen den latenten Variablen. Einfachpfeile repräsentieren direkte Effekte. So ist die Beziehung fw $\rightarrow$ int zu lesen als „Fachwissen erklärt (bzw. wirkt auf) die Verwendung der Argumentkategorie Intuition“ (siehe z. B. Eid, Gollwitzer & Schmitt, 2013, S. 606).

Tabelle 13: Fit-Indizes für das Modell `strukmodell`

Modell	χ^2	df	p	χ^2/df	CFI	RMSEA [95 % C.I.]; p	SRMR
`strukmodell`	1299.5	670	$< .01$	1.94	.94	.03 [.03; .04]; 1	.05

fekte quantifizieren nun die durch die exogene Variable bedingten Änderungen in den endogenen Variablen, bei gleichzeitiger Kontrolle bzw. Konstanthaltung aller anderen exogenen Variablen. Anders ausgedrückt handelt es sich bei dem direkten Effekt um die Steigung der Tangente der Funktion, die den (linearen) Zusammenhang zwischen endogener und exogener Variable beschreibt. Bei linearen Effekten kann die Stärke des Einflusses einer exogenen auf eine endogene Variable (bei gleichzeitiger Kontrolle aller anderen exogenen Variablen) durch eine einzige Zahl ausgedrückt werden. Dieser Pfadkoeffizient kann in seiner unstandardisierten bzw. standardisierten Form genau wie ein Regressionskoeffizient interpretiert werden (Kline, 2016, S. 232).

Die Fit-Indizes dieses Modells weisen eine akzeptable bis gute Passung auf und sind in Tabelle 13 zusammengefasst. Da die 16 Korrelationen zwischen den exogenen und endogenen Variablen aus dem globalen Messmodell des ersten Schrittes durch eine gleiche Anzahl an Regressionspfaden ersetzt werden, ändert sich die Zahl der freien Parameter nicht. Das modifizierte Modell weist im Vergleich zum ursprünglichen Messmodell daher eine identische Zahl an Freiheitsgraden auf; die Fit-Indizes ändern sich entsprechend nicht (zur Äquivalenz unterschiedlicher Strukturmodelle siehe z. B. Kline, 2016, S. 348 oder Urban und Mayerl, 2014, S. 46).

Die unstandardisierten sowie die vollständig standardisierten ML-Schätzer für die Regressionskoeffizienten sowie der Anteil aufgeklärter Varianz in den abhängigen Variablen R^2 sind in Tabelle 14 dargestellt. Die unstandardisierten Regressionskoeffizienten beschreiben den Zusammenhang in der Ursprungsmetrik der jeweiligen Variablen. Die standardisierten Regressionskoeffizienten (in der klassischen Regressionsanalyse mit manifesten Variablen oftmals als β-Gewichte oder Standardpartialregressionskoeffizienten bezeichnet) erlauben eine Aussage über den relativen Beitrag einer exogenen Variablen zur Varianzaufklärung innerhalb eines Modells. Auch wenn dieses Vorgehen umstritten ist, können standardisierte Regressionskoeffizienten verschiedener exogener Variablen innerhalb eines Modells verglichen werden (zur Diskussion bzgl. der Vergleichbarkeit von un- bzw. standardisierten Regressionskoeffizienten siehe z. B. Urban und Mayerl, 2014, S. 48 und Eid et al., 2013, S. 613). Eine Aufstellung der gängi-

gen Intervalle zur Interpretation der Effekte ist in Abschnitt E.2 dargestellt. Exemplarisch sei die „Lesart" von unstandardisiertem und standardisiertem Regressionskoeffizienten im folgenden Beispiel anhand des stärksten Regressors im Strukturmodell dargelegt; in der weiteren Ergebnisdarstellung wird im Fließtext nur noch der standardisierte Koeffizient berichtet, da die unstandardisierten Koeffizienten aufgrund der geringen inhaltlichen Bedeutung der hier überwiegend eingesetzten Likert-Skalen in der vorliegenden Arbeit nicht näher interpretiert werden.

Den insgesamt stärksten signifikanten Effekt nimmt das Fachwissen Mechanik auf die Verwendung der Kategorie Intuition. Der unstandardisierte Regressionskoeffizient beträgt $B_{\text{fw}\to\text{int}} = -0.49$, d. h. mit jeder Erhöhung des Fachwissens Mechanik um einen logit verringert sich die Verwendung der Argumentkategorie Intuition um -0.49 Einheiten auf der Likert-Skala. Der standardisierte Regressionskoeffizient beträgt $\beta_{\text{fw}\to\text{int}} = -.34$, d. h. dass bei Erhöhung des Fachwissens Mechanik um eine Standardabweichung die Verwendung der Argumentkategorie Intuition um $-.34$ Standardabweichungen abnimmt. Dieses Regressionsgewicht gibt an, inwieweit Variation in der Argumentkategorie Intuition auf den Prädiktor Fachwissen zurückzuführen ist, bei gleichzeitiger Konstanthaltung aller anderen exogenen Variablen (situationales Interesse, Kognitionsbedürfnis, Werteinschätzung der Naturwissenschaften). Je höher also das Fachwissen ausgeprägt ist, desto niedriger ist die Verwendung dieser Kategorie. Es handelt sich dabei um einen mittleren Effekt. Die Hypothese H1.1, die einen negativen Zusammenhang zwischen dem Fachwissen Mechanik und der Stärke der Verwendung der Kategorie Intuition postuliert, entspricht diesen Daten und wird daher angenommen. Die zugehörige Nullhypothese muss verworfen werden.

Einen weiteren signifikanten Einfluss hat das Fachwissen Mechanik auf die Verwendung der Kategorie Daten als Evidenz bei der Entscheidung zum Beibehalten oder Wechseln der eingangs aufgestellten Hypothese mit einem kleinen Effekt ($\beta_{\text{fw}\to\text{evi}} = .21$). Entsprechend wird die Hypothese H1.4 angenommen. Die Einflüsse des Fachwissens auf die Verwendung der Kategorie Messunsicherheiten (explizit) und Expertenwissen erreichen keine Signifikanz, die Hypothesen H1.2 und H1.3 müssen daher zugunsten der jeweiligen Nullhypothesen verworfen werden.

Die Ergebnisse der Strukturanalyse zeigen weiterhin, dass sowohl das Kognitionsbedürfnis als auch das situationale Interesse keinen signifikanten Beitrag zur Varianzaufklärung in allen vier Argumentkategorien liefern. In Bezug auf den Einfluss des Kognitionsbedürfnisses müssen daher die Hypothesen H1.5, H1.6, H1.7 und H1.8 zugunsten der jeweiligen Nullhypothese verworfen werden. In Bezug auf den Einfluss des situationalen Interesses auf die Verwendung der Argumentkategorien müssen daher die Hypothesen H1.13, H1.14, H1.15 und H1.16 verworfen werden.

Tabelle 14: Strukturparameter des Strukturgleichungsmodells `strukmodell`

Prädiktor:	fw →			nfc→			si→			wdn→			
Arg.-Kat.	B	SE	β	B	SE	β	B	SE	β	B	SE	β	R^2
int	−0.49*	(0.10)	−.34	−0.15	(0.10)	−.15	−0.07	(0.06)	−.07	0.32	(0.20)	.14	.15
exp	0.09	(0.11)	.06	−0.14	(0.11)	−.13	0.05	(0.07)	.05	0.52*	(0.23)	.22	.04
mu	−0.01	(0.06)	−.01	−0.01	(0.06)	−.02	0.01	(0.04)	.01	0.43*	(0.16)	.31	.09
evi	0.30*	(0.10)	.21	0.00	(0.09)	0.00	0.08	(0.06)	.08	0.12	(0.21)	.06	.08

Hinweise: * $\dots p < 0.05$, $B \dots$ unstandardisierter Regressionskoeffizient, $SE \dots$ Standardfehler, $R^2 \dots$ aufgeklärte Varianz der abhängigen Variablen. Die Irrtumswahrscheinlichkeiten sind der Übersicht halber nicht angegeben. Eine Signifikanztestung kann aber über die z-Statistik einfach nachvollzogen werden: Der Standardfehler des unstandardisierten Koeffizienten des Effekts von Fachwissen auf die Kategorie Intuition ist $SE = 0.10$, so dass $z = B/SE = -0.49/0.10 = -4.9$ über dem kritischen Wert von $z_{\text{krit},\alpha=0.05} = 1.96$ liegt (bei einem Signifikanzniveau von 5 %.

Die Skala Werteinschätzung der Naturwissenschaften (persönliche Relevanz) hat zunächst einen mittleren positiven Effekt auf die Verwendung der Kategorie Messunsicherheiten (explizit) ($\beta_{\text{wdn}\rightarrow\text{mu}} = .31$). Dies korrespondiert mit der Hypothese H1.11, die entsprechend angenommen wird. Weiterhin zeigt die persönliche Relevanz einen kleinen positiven Effekt auf die Verwendung der Argumentkategorien Expertenwissen ($\beta_{\text{wdn}\rightarrow\text{exp}} = .22$). Dies steht in Widerspruch zur Hypothese H1.10, die entsprechend zugunsten der Nullhypothese verworfen wird. Da die Daten dafür sprechen, dass hier die zu H1.10 entgegengesetzte Alternativhypothese gilt, wird diese hier angenommen. Die Hypothesen H1.9 und H1.12 werden verworfen, da keine signifikanten Effekte der persönlichen Relevanz auf die Kategorien Intuition und Daten als Evidenz vorliegen. Eine Zusammenfassung über die Bewertung aller Hypothesen findet sich am Ende des Ergebnisteils Abschnitt 5.6.

Anhand des multiplen Determinationskoeffizienten R^2 kann die Varianzaufklärung in den endogenen Variablen abgelesen werden. Die exogenen Variablen erklären 15 % der Varianz in der Argumentkategorie Intuition, 4 % der Varianz in der Argumentkategorie Expertenwissen, 9 % der Varianz in der Kategorie Messunsicherheiten (explizit) und 8 % der Varianz in der Argumentkategorie Daten als Evidenz.

Die Analysen dieses Strukturgleichungsmodells ergeben, dass eine Reihe von Pfaden keinen signifikanten Einfluss auf die Verwendung bestimmter Argumente hat (fw $\rightarrow$ exp + mu, si + nfc $\rightarrow$ int + exp + mu + evi, wdn $\rightarrow$ int+ evi). Zum Umgang mit nicht-signifikanten Pfaden existieren in der Literatur verschiedene Perspektiven. Kline (2016) diskutiert unter dem Stichwort „model trimming“ (S. 281) zwei verschiedene Ansätze. Lässt sich aus der Theorie begründen, dass ein Pfad nicht existiert und zu keiner signifikanten Modellverschlechterung führt, bedeutet dies einen Erkenntnisgewinn. Bei einer empirischen Respezifikation von Strukturgleichungsmodellen wird hingegen auf der Basis der Daten ein möglichst sparsames Modell entwickelt. Dabei wird meist kein signifikant besserer Modellfit erreicht und es besteht die Gefahr einer empirischen Anpassung des Modells an die vorliegende Stichprobe (Kline, 2016). In der vorliegenden Arbeit wurden daher die nicht-signifikanten Pfade im Modell belassen.

Anmerkung: Das Entfernen der nicht-signifikanten Pfade führt nicht zu bedeutsam veränderten Parameterschätzern und bringt damit auch keinen Vorteil im Hinblick auf die aufgeklärte Varianz in den endogenen Variablen mit sich. Die Modellschätzung bleibt stabil. Alle weiteren Analysen wurden daher auf Grundlage dieses Modells vorgenommen.

5.4 ANALYSE DER UNTERSCHIEDE IN DER VERWENDUNG DER ARGUMENTKATEGORIEN BEIM EXPERIMENTIEREN MIT REALEXPERIMENT BZW. COMPUTERSIMULATION

Die Analysen im nachfolgenden Abschnitt adressieren die Gruppenunterschiede bzgl. der Stärke der Verwendung bestimmter Argumentkategorien zwischen den Probanden, die mit dem Realexperiment bzw. der Computersimulation gearbeitet haben (vgl. Forschungsfrage 3). Die Unterschiedshypothesen wurden im Rahmen der Strukturgleichunsmodellierung durch die Verwendung von Multigruppen-Strukturgleichungsmodellen MG-SGM getestet. Eine methodische Einführung in die Verwendung von MG-SGM findet sich z. B. bei Urban und Mayerl (2014, S. 229), eine tiefergehende Darstellung des Verfahrens findet sich bei Green und Thompson (2012, S. 393) sowie Kline (2016, S. 394). Grundlegende Voraussetzung für das Testen von Mittelwertsunterschieden ist die Äquivalenz der Messmodelle über verschiedene Teilstichproben hinweg. Die Messinvarianz des verwendeten Tests wurde bereits während der Testentwicklung evaluiert (vgl. Anhang B) als auch anhand dieser Stichprobe evaluiert (vgl. Abschnitt 4.9.7). Da die Äquivalenz der Messmodelle bestätigt wurde, können nun die Unterschiedshypothesen H3.4, H3.2, H3.1, H3.3 getestet werden, d. h. es kann untersucht werden, ob sich die Mittelwerte der Argumentkategorien signifikant über die Gruppen hinweg unterscheiden. Zur Identifizierung der latenten Mittelwertstruktur wurde die sog. Referenzgruppen-Methode herangezogen (Kline, 2016, S. 404). Dabei werden die latenten Mittelwerte der Faktoren in einer Gruppe auf null fixiert. Diese Gruppe dient somit als Referenzgruppe. In der anderen Gruppe werden dann die latenten Mittelwerte und Faktorvarianzen frei geschätzt. In der vorliegenden Arbeit wurde als Referenzgruppe die Gruppe der Probanden, die mit dem Realexperiment gearbeitet haben, festgelegt. Die so geschätzten latenten Mittelwerte in der Gruppe der Probanden, die mit dem Computerexperiment gearbeitet haben, können dann als Differenzen zur ersten Gruppe interpretiert werden. Als Metrik für die unstandardisierten Schätzer wird die Originalmetrik verwendet, hier also die Metrik der fünfstufigen Likert-Skala. Die durch `lavaan` ausgegebenen standardisierten latenten Mittelwertsunterschiede können entsprechend der *d*-Statistik nach Cohen (1988) und den gängigen Intervallen zur Einschätzung der Effektstärke interpretiert werden (vgl. Abschnitt E.2).

5.4.1 *Analyse der Gruppenunterschiede ohne Kontrolle des Einflusses der personalen Faktoren*

Zur Analyse der Mittelwertsunterschiede in der Stärke der Verwendung der vier Argumentkategorien wurde in einem ersten Schritt ein Multigruppen-Strukturgleichungsmodell geschätzt, *ohne* den Einfluss

Tabelle 15: Fit-Indizes für das Multigruppen-Strukturgleichungsmodell, das zur Evaluation der Mittelwertsunterschiede herangezogen wurde (ohne Kontrolle der personalen Faktoren)

	χ^2	df	p	χ^2/df	CFI	RMSEA [95 % C.I.]; p	SRMR
`arg-4fak-mw`	641.3	352	$< .01$	1.82	.94	.05 [.04; .05]; .96	.07

der personalen Faktoren auf die Stärke der Verwendung bestimmter Argumente zu berücksichtigen. Dieses Vorgehen verhält sich methodisch analog zu einer einfaktoriellen multivariaten Varianzanalyse (MANOVA) mit dem Faktor Gruppenzugehörigkeit, überwindet jedoch einige Nachteile dieser (für einen Vergleich von MANOVA und der Analyse von Mittelwerten in Strukturgleichungsmodellen siehe z. B. Green & Thompson, 2012).

Es wurde dafür zunächst ein vierfaktorielles Modell spezifiziert, das die Struktur des Argumentationstests abbildet. Die Faktorstruktur, -ladungen und Item-Intercepts wurden in beiden Gruppen auf einen identischen Wert fixiert, so dass ein Modell skalarer Invarianz vorliegt. In diesem Modell werden die Mittelwerte der Argumentkategorien in der Referenzgruppe auf null fixiert; die latenten Faktorvarianzen frei geschätzt. In der Vergleichsgruppe (hier: die Gruppe der Probanden, die mit dem Computerexperiment gearbeitet haben) werden die latenten Mittelwerte und die latenten Faktorvarianzen frei geschätzt. Die Fit-Indizes für dieses Modell sind in Tabelle 15 zusammengefasst. Das Modell weist eine gute Passung zu den Daten auf.

Die Parameterwerte der ML-Schätzung sind in Tabelle 16 dargestellt. Unstandardisierte Mittelwerte und Varianzen liegen in der Metrik der Skala vor, die standardisierten Parameter können wie ein standardisierter Mittelwertsunterschied, d. h. wie die d-Statistik nach Cohen (1988) interpretiert werden. Dabei kann konzeptuell zwischen verschiedenen Varianten unterschieden werden. Eine Erläuterung dazu findet sich in Abschnitt E.2.

Entsprechend der verwendeten Referenzgruppen-Methode (siehe oben) sind die Mittelwerte der Gruppe der Probanden, die mit dem Realexperiment gearbeitet haben, zu null fixiert. In der anderen Gruppe (Computersimulation) wurden die Mittelwerte frei geschätzt. Die Signifikanz des Unterschiedes zwischen den Mittelwerten wurde über die Welch-James-t-Statistik bestimmt (Kline, 2016, S. 410). Die Kategorie Intuition wurde signifikant weniger in der Gruppe V (Computersimulation) verwendet. Der Unterschied entspricht mit -0.16 Einheiten auf der Likert-Skala einem standardisierten Mittelwertsunterschied von $d_{\text{int,V}} = 0.21$, der nicht den Schwellenwert von $d \geq 0.3$ für einen

kleinen Effekt erreicht. Argumente aus der Kategorie Expertenwissen wurden in beiden Gruppen gleich häufig verwendet. Das Berücksichtigen von Messunsicherheiten in der Argumentation trat in der Gruppe V signifikant weniger auf als in der Gruppe R, wobei es sich mit $d_{\mathrm{mu,V}} = -1.14$ um einen großen Effekt handelt. Ebenfalls signifikant ist der Unterschied in der Verwendung der Kategorie Daten als Evidenz, der jedoch mit $d_{\mathrm{evi,V}} = 0.14$ ebenfalls derart klein ist, dass er vernachlässigt werden kann (zur Interpretation von standardisierten Mittelwertsunterschieden siehe Abschnitt E.2).

Die berichteten Mittelwertsunterschiede lassen sich auch in den Histogrammen der Skalen wiederfinden, die durch Summierung der jeweiligen Antworten auf der Likert-Skala erhalten werden können (vgl. Abschnitt C.5.1).

Tabelle 16: Unstandardisierte und standardisierte ML-Schätzer für die Mittelwertstruktur des Argumentationstests für die Gruppen der Probanden, die mit dem Realexperiment (Gruppe R) bzw. dem Computerexperiment (Gruppe V) experimentiert haben.

	Gruppe R ($n = 427$)			Gruppe V ($n = 394$)			
Parameter	Unst.	SE	Std.	Unst.	SE	Std.	p
Intuition							
Varianz	0.59	0.06	1	0.64	0.06	1	
Mittelwert	0.00		0.00	−0.16	0.06	−0.21	.006
Expertenwissen							
Varianz	0.61	0.07	1	0.82	0.10	1	
Mittelwert	0.00		0.00	0.09	0.07	0.10	.193
Messunsicherheiten							
Varianz	0.15	0.04	1	0.22	0.05	1	
Mittelwert	0.00		0.00	−0.53	0.06	−1.14	$< .001$
Daten als Evidenz							
Varianz	0.48	0.05	1	0.74	0.09	1	
Mittelwert	0.00		0.00	0.12	0.06	0.14	.045

Hinweise: Unst. ... unstandardisierter Parameter; Std. ... Standardisierter Parameter; SE ... Standardfehler; p ... Irrtumswahrscheinlichkeit (Welch-James-t-Statistik). Standardisierte Mittelwertparameter können wie die d-Statistik nach Cohen (1988) interpretiert werden (vgl. Abschnitt E.2).

5.4.2 *Analyse der Gruppenunterschiede in der Verwendung der Argumentkategorien mit Kontrolle der personalen Faktoren*

Im vorherigen Abschnitt wurden die Gruppenunterschiede in der Stärke der Verwendung der vier Argumentkategorien ohne Kontrolle von individuellen Unterschieden in den personalen Faktoren berechnet. Im Folgenden wird die Mittelwertstruktur der Argumentkategorien nun unter Berücksichtigung des in Abschnitt 5.3 dargestellten Modells analysiert. Dabei werden die vier latenten Faktoren, welche die Argumentkategorien darstellen, auf die personalen Faktoren regreggiert. Durch dieses Vorgehen wird die Mittelwertstruktur um den Einfluss der personalen Faktoren korrigiert. Dieses Vorgehen ist dabei konzeptuell mit einer multivariaten Kovarianzanalyse (MANCOVA) zu vergleichen (Green & Thompson, 2012, S. 404), die natürlich die Vorteile der Strukturgleichungsmodellierung nicht aufweist. Durch dieses Vorgehen wird die Mittelwertstruktur der abhängigen Variablen um die Varianz, die durch die personalen Faktoren erzeugt wird, korrigiert (z. B. Eid et al., 2013). Ausgehend von einem MG-SGM mit metrischen Invarianzbedingungen (d. h. Faktorstruktur, Faktorladungen und Item-Intercepts sind in beiden Gruppen auf einen identischen Wert fixiert, vgl. Abschnitt B.7.2.2) wurde in einem ersten Schritt ein Modell geschätzt, das eine Variation der Regressionskoeffizienten über die Gruppen hinweg zulässt (Modell `strukmod-mg` in Tabelle 17). In einem zweiten Modell wurden dann die 16 unstandardisierten Regressionskoeffizienten in beiden Gruppen auf einen gleichen Wert beschränkt (Modell `strukmod-mg` in Tabelle 17). Durch einen Vergleich des Modellfits kann dann überprüft werden, ob sich die Wirkzusammenhänge zwischen den personalen Faktoren und den Argumentkategorien über die Gruppen hinweg unterscheiden. Dieses Vorgehen ist vergleichbar mit der Überprüfung der Voraussetzung der Homogenität der Regressionskoeffizienten (*homogenity of regression slopes*, siehe z. B. Field et al., 2012, oder Eid et al., 2013, S. 722) in der klassischen Kovarianzanalyse (ANCOVA). Würden sich die Steigungen der Regressionsgeraden über die Gruppen hinweg unterscheiden, würde dies auf eine Interaktion zwischen Gruppenzugehörigkeit und personalen Faktoren hindeuten.

Die Fit-Indizes für beide Modelle sind in Tabelle 17 dargestellt. Beide Modelle weisen eine relativ gute Passung zu den Daten auf. Der Modellvergleich durch einen χ^2-*likelihood-ratio*-Test zeigt keine signifikante Verschlechterung des Modellfits an ($\Delta\chi^2(16) = 23.06$, $p = .11$). Es kann daher die Hypothese der Homogenität der Regressionskoeffizienten nicht verworfen werden, so dass die Ergebnisse sinnvoll interpretiert werden können.

Die Parameter der ML-Schätzung dieses Modells sind in Tabelle 18 dargestellt. Neben den latenten Mittelwerten sind auch die unstandardisierten und standardisierten Regressionskoeffizienten für die di-

Tabelle 17: Fit-Indizes für die Modelle zur Überprüfung der Homogenität der Regressionskoeffizienten

Modell	χ^2	df	p	χ^2/df	CFI	RMSEA [95 % C.I.]; p	SRMR
`strukmod-mg`	2160.8	1402	< 0.01	1.54	0.929	0.04 [0.033; 0.039]; 1	0.06
`strukmod-mg-hmg`	2183.9	1418	< 0.01	1.54	0.928	0.04 [0.033; 0.039]; 1	0.06

rekten Effekte der personalen Faktoren auf die Verwendung der Argumentkategorien dargestellt. Wie im vorhergehenden Modell sind die Mittelwerte der Gruppe der Probanden, die mit dem Realexperiment gearbeitet haben, zu null fixiert. Die unstandardisierten Parameter der Regressionskoeffizienten sind über die Gruppen hinweg auf einen gleichen Wert fixiert. Da sich die latenten Faktorvarianzen aber weiterhin unterscheiden können, sind die standardisierten Effekte über die Gruppen hinweg marginal unterschiedlich (vgl. Urban und Mayerl, 2014, S. 223 bzw. Kline, 2016, S. 395). Der Vergleich der Regressionskoeffizienten aus diesem Modell mit den im vorhergehenden Abschnitt berichteten Koeffizienten aus dem Modell ohne Gruppenvergleich zeigt, dass sich die unstandardisierten Parameter für die Effekte der personalen Faktoren auf die Argumentkategorien nicht bedeutsam unterscheiden (vgl. Tabelle 14). Zur Verdeutlichung sei hier ein Beispiel angegeben: So ist der signifikante negative Effekt des Fachwissens auf die Verwendung der Argumentkategorie Intuition im Modell ohne Gruppenvergleich $B_{\text{fw}} \rightarrow \text{evi} = 0.30$, im Modell mit Gruppenvergleich wird dieser Regressionskoeffizient zu $B_{\text{fw}} \rightarrow \text{evi} = 0.28$ bestimmt.

In Bezug auf die Mittelwertsunterschiede zeigen sich auch in diesem Modell, welches den Einfluss der personalen Faktoren in der Verwendung der Argumentkategorien kontrolliert, drei signifikante Mittelwertsunterschiede in der Stärke der Verwendung der Argumentkategorien Intuition ($d_{\text{int,V}} = 0.26$), Messunsicherheiten ($d_{\text{mu,V}} = -1.08$) sowie Daten als Evidenz ($d_{\text{evi,V}} = 0.18$). Diese Effekte sind in ihrer Ausprägung ebenfalls vergleichbar mit dem Modell ohne Kontrolle der personalen Faktoren (vgl. Tabelle 16).

Tabelle 18: Unstandardisierte und standardisierte ML-Schätzer für die Mittelwertsstruktur des Argumentationstests für die Gruppen der Probanden, die mit dem Realexperiment (Gruppe R) bzw. dem Computerexperiment (Gruppe V) experimentiert haben, sowie die direkten Effekte der personalen Faktoren auf die Argumentkategorien je Gruppe (Parameter entnommen aus Modell `strukmod-mg-hmg`, vgl. Tabelle 17)

	Gruppe R ($n = 427$)				Gruppe V ($n = 394$)				
	Unst.	*SE*	Std.	R^2	Unst.	*SE*	Std.	R^2	p
Intuition				.18				.14	
Varianz	0.47	0.06	1		0.58	0.06	1		
Mittelwert	0.00		.00		−0.21	0.06	−.26		< 0.00
fw → int	−0.49	0.10	−.36		−0.49	0.10	−.31		.00
nfc → int	−0.19	0.10	−.18		−0.19	0.10	−.18		.06
si → int	−0.07	0.06	−.07		−0.07	0.06	−.07		.28
wdn → int	0.35	0.20	.16		0.35	0.20	.16		.08
Expertenwissen				.04				.03	
Varianz	0.59	0.07	1		0.79	0.10	1		
Mittelwert	0.00		.00		0.11	0.07	.12		.11
fw → exp	0.04	0.11	.03		0.04	0.11	.03		.68
nfc → exp	−0.10	0.11	−.09		−0.10	0.11	−.09		.36
si → exp	0.05	0.07	.05		0.05	0.07	.04		.50
wdn → exp	0.48	0.23	.22		0.48	0.23	.19		.03
Messunsicherheiten				.14				.09	
Varianz	0.13	0.03	1		0.21	0.05	1		
Mittelwert	0.00		.00		−0.52	0.06	−1.08		< 0.00
fw → mu	0.02	0.06	.02		0.02	0.06	.02		.80
nfc → mu	−0.06	0.06	−.11		−0.06	0.06	−.09		.33
si → mu	0.01	0.04	.02		0.01	0.04	.02		.78
wdn → mu	0.46	0.14	.42		0.46	0.14	.35		0.00
Daten als Evidenz				.10				.06	
Varianz	0.43	0.05	1		0.70	0.09	1		
Mittelwert	0.00		.00		0.15	0.06	.18		.01
fw → evi	0.28	0.09	.23		0.28	0.09	.17		0.00
nfc → evi	0.05	0.09	.05		0.05	0.09	.04		.62
si → evi	0.07	0.06	.07		0.07	0.06	.06		.28
wdn → evi	0.08	0.20	.04		0.08	0.20	.03		.71

Hinweise: Unst. ... unstandardisierter Parameter; Std. ... Standardisierter Parameter; *SE* ... Standardfehler; R^2 aufgeklärte Varianz in den abhängigen Variablen; p ... Irrtumswahrscheinlichkeit (Welch-James-*t*-Statistik). Standardisierte Mittelwertparameter können im Sinne der *d*-Statistik nach Cohen (1988) interpretiert werden (vgl. Abschnitt E.2).

5.4.3 *Bewertung der Unterschiedshypothesen*

Auf Grundlage der berichteten Modelle sollen nun die statistischen Hypothesen hinsichtlich der Gruppenunterschiede in der Stärke der Verwendung der Argumentkategorien ausgewertet werden. Diese Analyse adressiert Forschungsfrage 3. Die Mittelwertvergleiche aus beiden Modellen, d. h. mit und ohne Kontrolle der personalen Faktoren, lassen identische Schlussfolgerungen zu. Es sei an dieser Stelle daran erinnert, dass es sich bei diesen Hypothesen um gerichtete Alternativhypothesen handelt, die dem theoretisch begründbaren Effekt entsprechen. Die entsprechenden Nullhypothesen negieren den angenommenen Effekt (vgl. Abschnitt 3.2).

Im Hinblick auf den Unterschied in der Stärke der Verwendung der Argumentkategorie Intuition verwenden Probanden, die mit dem Computer gearbeitet haben, diese Kategorie signifikant weniger häufig. Es handelt sich mit $d_{\text{int, V}} = -0.26$ um einen geringen Effekt, der nicht das Kriterium für einen kleinen Effekt erreicht (zur Interpretation von Effektstärken siehe Abschnitt E.2). Vor dem Hintergrund dieser Daten muss daher die Hypothese H3.1 zugunsten der Nullhypothese verworfen werden. Die Richtung des beobachteten Effekts spricht hier für einen zu H3.1 gegenläufigen Effekt, es wird daher die entsprechende Alternativhypothese angenommen. Im Hinblick auf die Stärke der Verwendung von Argumenten aus der Kategorie Expertenwissen zeigt sich kein signifikanter Unterschied zwischen den beiden Gruppen. Daher wird die Nullhypothese zu H3.2 verworfen und die Forschungshypothese angenommen. Die Stärke der Kategorie Messunsicherheiten (explizit) wurde in der Gruppe der Probanden, die mit dem Computerexperiment gearbeitet haben, signifikant seltener verwendet. Es handelt sich dabei mit $d_{\text{mu, V}} = -1.08$ um einen großen Effekt. Dieses Ergebnis entspricht der Hypothese H3.3, die daher angenommen wird. Die Kategorie Daten als Evidenz wird in der Gruppe der Probanden, die mit dem Computerexperiment gearbeitet haben, signifikant häufiger verwendet. Aus diesem Grund muss die Hypothese H3.4 hier verworfen werden. Stattdessen wird die entgegengesetzte Alternativhypothese angenommen. Wobei anzumerken ist, dass es sich hierbei um einen minimalen Effekt handelt, der nicht interpretierbar ist.

5.5 ANALYSE DES EINFLUSSES DER ARGUMENTKATEGORIEN AUF DEN LERNERFOLG

5.5.1 *Der Einfluss der Verwendung bestimmter Argumente auf die Richtigkeit der nach dem Experiment aufgestellten Hypothese*

Die Forschungsfrage 2 untersucht den Einfluss der Stärke der Verwendung bestimmter Argumente auf die Richtigkeit der nach dem

Experiment aufgestellten Hypothese. Der folgende Abschnitt analysiert die Daten im Hinblick auf diese Fragestellung. Wie Abbildung 6 zu entnehmen ist, behalten von den 821 Probanden, die eingangs eine falsche Hypothese aufgestellt haben, 296 Probanden (36 %) diese fachlich inkorrekte Hypothese nach dem Experimentieren bei. 520 Probanden (63 %) stellen eine neue Hypothese auf, davon entscheiden sich 424 der Probanden (82 %) für eine fachlich adäquate Hypothese, während 94 (18 %) nach dem Experiment von einer falschen Hypothese zu einer anderen falschen Hypothese wechseln. Nach dem Experimentieren stellen also 390 Probanden eine falsche Hypothese auf, 494 Probanden stellen eine fachlich adäquate Hypothese auf.

Die nach dem Experiment aufgestellte Hypothese kann binär kodiert werden (0 ... falsch, 1 ... richtig, vgl. Abschnitt 4.3.2). Es ist daher zunächst eine Möglichkeit, anhand dieser Variablen die Stichprobe in zwei Gruppen zu splitten und die Mittelwerte der Verwendung der Argumentkategorien auf Gleichheit zu testen. Da dieses Verfahren Schwächen aufweist, sind die Ergebnisse dieses Gruppenvergleichs nur zum Zweck der Orientierung in Abschnitt D.1 dargestellt. Denn zum einen berücksichtigt dieses Verfahren methodisch nicht die simultan wirkenden Einflüsse der Argumentkategorien auf die Entscheidung zum Aufstellen der Hypothese sowie die Zusammenhänge zwischen den Argumentkategorien (vgl. Tabelle 12). Zum anderen liegt einem solchen Gruppenvergleich die Annahme zugrunde, dass zwei Grundgesamtheiten existieren (Lernende, die nach diesem Experiment eine richtige bzw. eine falsche Hypothese aufstellen), aus denen je eine Zufallsstichprobe gezogen wurde. Diese Annahme wird jedoch im vorliegenden Kontext verletzt, da das Aufstellen der Hypothese durch die Situation bedingt wurde. Diese Argumentation trifft beim Gruppenvergleich – reales Experiment vs. Computerexperiment – nicht zu, da hier die Stichprobe a priori randomisiert den beiden Gruppen zugeordnet wurde.

Ein geeigneteres Verfahren zur Untersuchung des Einflusses der Argumentkategorie auf die Richtigkeit der aufgestellten Hypothese nach dem Experiment stellen multiple logistische Regressionsmodelle dar. Logit-Modelle sind dazu geeignet, Einflüsse verschiedenster Faktoren auf eine diskrete, z. B. binäre bzw. kategoriale abhängige Variable zu untersuchen (z. B. Döring & Bortz, 2016, S. 678). In Abschnitt E.3 sind Anmerkungen zur logistischen Regression aufgeführt. Tiefergehende Darstellungen sind z. B. in Behnke (2015), Hosmer, Lemeshow und Sturdivant (2013), Urban (1993) zu finden. Im Gegensatz zur linearen Regression wird in binären Logit-Modellen nicht die Ausprägung der abhängigen Variablen, sondern die Wahrscheinlichkeit für das Eintreten eines Ereignisses, das mit der abhängigen Variablen korrespondiert (hier: die fachlich adäquate Hypothese nach dem Experiment, kodiert mit eins), modelliert.

In einem ersten Schritt wurden vier Modelle spezifiziert, in denen die vier Argumentkategorien zunächst nur einzeln als Prädiktor auf die binäre abhängige Variable in Form der Richtigkeit der Hypothese nach dem Experiment wirken. Die Parameter dieser Modelle sind in Tabelle 19 dargestellt. Im Gegensatz zu den linearen Strukturgleichungsmodellen in den vorhergehenden Abschnitten können bei Strukturgleichungsmodellen, die logistische Regressionen enthalten, derzeit keine relativen Modellfit-Indizes angegeben werden (B. O. Muthén, 1998–2004; B. O. Muthén & L. K. Muthén, 1998–2015).

Es zeigt sich, dass alle vier Argumentkategorien signifikant die Wahrscheinlichkeit für das Aufstellen der richtigen Hypothese nach dem Experimentieren erklären. Die Stärke der Verwendung der Kategorie Intuition hat einen negativen Effekt auf die Richtigkeit der Hypothese (Modell `logreg-3` in Tabelle 19, $\beta_{\text{int}} = -0.54, p < .001$). Durch Exponentieren des Regressionskoeffizienten β kann das *odds ratio* (OR, Chancenverhältnis) erhalten werden. Das Chancenverhältnis ergibt sich zu $OR_{\text{int,post}} = \exp(-0.54) = 0.58$. Das bedeutet, dass sich bei Erhöhung der Stärke der Verwendung dieser Kategorie um eine Einheit auf der Likert-Skala die Chance auf eine fachlich adäquate Hypothese um den Faktor $OR_{\text{int,post}} = 0.58$ verringert. Diese Aussage ist äquivalent zu der Aussage, dass sich die Chance auf eine falsche Hypothese um den Faktor $\frac{1}{0.58} = 1.7$ erhöht. Es ist daher von einem kleinen Effekt zu sprechen (zur Interpretation von *odds ratios* siehe Abschnitt E.2.4).

In logistischen Regressionsmodellen können R^2-Bestimmtheitsmaße nicht wie in linearen Modellen bestimmt werden. Dies liegt u. a. darin begründet, dass bei binären Variablen keine Varianz vorliegt. Daher existieren eine Reihe von sog. Pseudo-Bestimmtheitsmaßen. Hier wurde das Pseudo-Bestimmtheitsmaß nach McKelvey und Zavoina (1975) bestimmt. Dieses Maß kann als die erklärte Varianz, einer dem beobachteten Antwortverhalten $(0/1)$ zugrundeliegenden, normalverteilten latenten Variablen interpretiert werden. Der Pseudo-R^2 für das Modell liegt bei 5.1 %. Ebenfalls einen kleinen negativen Effekt hat die Verwendung der Kategorie Messunsicherheiten (explizit) (Modell `logreg-3` in Tabelle 19, $\beta_{\text{mu}} = -0.52, p < .001, OR_{\text{mu,post}} = 0.59, R^2 = 4.9\,\%$). Einen kleinen positiven Effekt auf die Richtigkeit der Hypothese nach dem Experiment hat die Verwendung der Kategorie Expertenwissen (`logreg-2`, $\beta_{\text{exp}} = 0.57, p < .001, OR_{\text{exp,post}} = 1.77, R^2 = 6.6\,\%$). Einen signifikanten positiven Einfluss auf die Richtigkeit der Hypothese hat die Verwendung der Argumentkategorie Daten als Evidenz (Modell `logreg-4`, $\beta_{\text{evi}} = 1.28, p < .001, OR_{\text{evi,post}} = 3.58$), hierbei ist von einem mittleren bis großen Effekt zu sprechen, was sich auch in der aufgeklärten Variabilität zeigt (Pseudo-$R^2 = 23.1\,\%$).

Um die Zusammenhänge zwischen den Prädiktoren zu berücksichtigen, wurde zudem ein multiples logistisches Regressionsmodell

spezifiziert, bei dem alle vier Argumentkategorien simultan als Prädiktoren zur Erklärung der Richtigkeit der Hypothese als abhängige Variable eingehen. Die Parameter dieses Modells sind in Tabelle 19 (Modell: `logreg-5`) dargestellt. Es zeigt sich, dass die Regressionskoeffizienten der Kategorien Intuition und Expertenwissen nun nicht mehr signifikant von null verschieden sind. Es zeigt sich ein signifikanter negativer Einfluss der Argumentkategorie Messunsicherheiten (explizit) ($\beta_{\text{mu}} = -0.80, p < .001, OR_{\text{mu,post}} = 0.45$). Das bedeutet, dass bei Erhöhung der Stärke der Verwendung der Kategorie Messunsicherheiten (explizit) um einen Punkt auf der Likert-Skala – und gleichzeitiger Konstanthaltung aller anderen Prädiktoren – die Chance auf eine fachlich adäquate Hypothese um den Faktor 0.45 sinkt (die Chance für eine falsche Hypothese steigt um $\frac{1}{0.45} = 2.22$), was etwa einem mittleren Effekt entspricht. Ein signifikanter positiver Einfluss zeigt sich bei der Kategorie Daten als Evidenz ($\beta_{\text{evi}} = 1.46, p < .001, OR_{\text{evi,post}} = 4.31$). Hier ist von einem großen Effekt zu sprechen. Das Pseudo-R^2 nach McKelvey und Zavoina (1975) beträgt für das multiple logistische Regressionsmodell 31.9 %.

Tabelle 19: Parameter der logistischen Regressionsmodelle zur Schätzung des Einflusses der Argumentkategorien auf die Richtigkeit der Hypothese nach dem Experiment

Modell	β	SE_β	p	OR (e^β) [95 % C.I.]	LL	BIC	R^2
`logreg-1`					−22 587.4	45 429.3	5.1 %
int	−0.54	0.11	< .001	0.58 [0.47; 0.73]			
`logreg-2`					−22 584.8	45 652.9	6.6 %
exp	0.57	0.12	< .001	1.77 [1.41; 2.21]			
`logreg-3`					−22 589.8	45 434.1	4.9 %
mu	−0.52	0.13	< .001	0.59 [0.46; 0.76]			
`logreg-4`					−22 537.0	45 328.5	23.1 %
evi	1.28	0.16	< .001	3.58 [2.64; 4.85]			
`logreg-5`					−22 516.7	45 536.9	31.9 %
int	0.20	0.16	.198	1.22 [0.90; 1.66]			
exp	0.10	0.14	.461	1.11 [0.85; 1.45]			
mu	−0.80	0.16	< .001	0.45 [0.33; 0.61]			
evi	1.46	0.22	< .001	4.31 [2.79; 6.65]			

Hinweise: $n = 821$, β ... unstandardisierter Regressionskoeffizient; SE_β ... Standardfehler von β; OR ... *odds ratio* (e^β), LL ... Log-Likelihood; BIC ... Bayesian Information Criterion; R^2 ... Pseudo-Bestimmtheitsmaß nach McKelvey und Zavoina (1975).

5.5.2 *Einfluss der Verwendung bestimmter Argumente auf die Richtigkeit der Hypothese zum Zeitpunkt der Follow-up-Erhebung*

Zur Analyse des Einflusses der Stärke der Verwendung der vier Argumentkategorien auf die Richtigkeit der Hypothese zum Zeitpunkt der Follow-up Erhebung wurden – analog zum vorhergehenden Abschnitt – zunächst vier Regressionsmodelle spezifiziert, in denen die Wahrscheinlichkeit für die Richtigkeit der Hypothese durch die Verwendung je einer Argumentkategorie erklärt wird (Modelle `logreg-fu1` bis `logreg-fu-4` in Tabelle 20). Das Modell, in dem die Kategorie Intuition die Richtigkeit der Hypothese zum Zeitpunkt der Follow-up-Erhebung erklärt, zeigt einen signifikanten negativen Einfluss dieser Kategorie (`logreg-fu-1`, $\beta_{\text{int,fu}} = -0.51, p < .001, OR_{\text{int,fu}} = 0.60$). Dieser Effekt ist als klein zu bewerten (vgl. Abschnitt E.2.4). Die Kategorie Daten als Evidenz weist ebenfalls einen kleinen positiven Effekt auf (Modell `logreg-fu-4`, $\beta_{\text{evi,fu}} = 0.46, p = .001, OR_{\text{evi,fu}} = 1.58$), während die Kategorien Messunsicherheiten (explizit) (Modell `logreg-fu-3` $\beta_{\text{mu,fu}} = -0.04, p = .48, OR_{\text{mu,fu}} = 0.91$) und Expertenwissen (Modell `logreg-fu-2` $\beta_{\text{exp,fu}} = 0.11, p = .25, OR_{\text{mu,fu}} = 1.14$) keinen signifikanten Einfluss auf die Nachhaltigkeit des Hypothesenwechsels zeigen.

Bei Überführung aller vier Argumentkategorien als Prädiktoren in ein multiples logistisches Regressionsmodell zur Erklärung der Richtigkeit der Hypothese zum Zeitpunkt der Follow-up-Erhebung erweist sich nur die Argumentkategorie Intuition mit einem kleinen negativen Effekt als signifikant (Modell `logreg-gu-5`, $\beta_{\text{int}} = -0.38, p < .009, OR_{\text{int,fu}} = 0.69$). Die Kategorien Expertenwissen, Messunsicherheiten (explizit) und Daten als Evidenz zeigen keinen signifikanten Einfluss zur Vorhersage der Wahrscheinlichkeit für die fachlich adäquate Hypothese zum Zeitpunkt der Follow-up Erhebung. Das Pseudo-R^2 für dieses Modell beträgt 5.1 %. Die Richtigkeit der Hypothese lässt sich unmittelbar nach dem Experiment also deutlich besser erklären, als zum Zeitpunkt der Follow-up-Erhebung. Bei dem hier dargestellten Vorgehen zur Analyse der Nachhaltigkeit des Lernerfolgs wird eine methodische Unsauberkeit in Kauf genommen: Durch die wiederholte Erhebung der Hypothese zum Follow-up-Zeitpunkt werden längsschnittliche Daten erzeugt. Damit liegt potentiell ein systematischer Zusammenhang zwischen (den Residuen) der abhängigen Variablen vor, dem durch das berichtete Vorgehen nicht Rechnung getragen wurde. Da über den Umgang mit längsschnittlichen binären Variablen zum einen Unklarheit besteht (Williamson, Bangdiwala, Marshall & Waller, 1996), zum anderen sehr komplexe Lösungen vorgeschlagen werden (Landerman, Mustillo & Land, 2011), wurde sich hier für die einfachste Möglichkeit entschieden. Konkret bedeutet dies, dass der Einfluss der Argumentkategorien auf die abhängige Variable in Form der fachlichen Richtigkeit zu den beiden Testzeitpunkten in zwei getrennten Modellen bestimmt wurde.

Tabelle 20: Parameter der logistischen Regressionsmodelle zur Schätzung des Einflusses der Argumentkategorien auf die Richtigkeit der Hypothese zum Zeitpunkt der Follow-up-Erhebung

Modell	β	SE_β	p	OR (e^β) [95 % C.I.]	LL	BIC	R^2
`logreg-fu-1`					−22 452.7	45 160.0	4.5 %
int	−0.51	0.12	< .001	0.60 [0.47; 0.76]			
`logreg-fu-2`					−22 461.7	45 178.0	0.4 %
exp	0.11	0.10	.253	1.14 [0.91; 1.43]			
`logreg-fu-3`					−22 462.2	45 179.0	0.1 %
mu	−0.04	0.05	.479	0.91 [0.72; 1.17]			
`logreg-fu-4`					−22 455.0	45 164.4	3.7 %
evi	0.46	0.14	.001	1.58 [1.21; 2.06]			
`logreg-fu-5`					−22 451.0	45 167.2	5.1 %
int	−0.38	0.14	.009	0.69 [0.52; 0.91]			
exp	0.02	0.13	.904	1.02 [0.79; 1.30]			
mu	−0.03	0.12	.829	0.97 [0.77; 1.24]			
evi	0.25	0.17	.138	1.28 [0.92; 1.77]			

Hinweise: $n = 640$, β ... unstandardisierter Regressionskoeffizient; SE_β ... Standardfehler von β; OR ... *odds ratio* (e^β), LL ... Log-Likelihood; BIC ... Bayesian Information Criterion; R^2 ... Pseudo-Bestimmtheitsmaß nach McKelvey und Zavoina (1975).

5.5.3 *Bewertung der statistischen Hypothesen zum Einfluss der Argumentkategorien auf den Lernerfolg*

Vor dem Hintergrund der Ergebnisse der logistischen Regressionsmodelle können nun die statistischen Hypothesen bewertet werden, die den Einfluss der Stärke der Verwendung der vier Argumentkategorien auf den Lernerfolg adressieren.

Dabei zeigt sich, dass je nach ausgewähltem Modell die Hypothesen unterschiedlich bewertet werden können. Wird z. B. im Hinblick auf den Einfluss der Kategorie Intuition auf die Richtigkeit der Hypothese nach dem Experimentieren das singuläre Regressionsmodell, das als Prädiktor nur die Kategorie Intuition enthält, herangezogen, zeigt sich ein kleiner negativer Effekt (Modell `logreg-1` in Tabelle 19). Die Hypothese H2.1 müsste daher angenommen werden. Wird in einem multiplen Regressionsmodell nun aber gleichzeitig die gemeinsame Varianz der Prädiktoren kontrolliert, ist dieser Effekt nicht mehr signifikant von null verschieden (Modell `logreg-5`). Dies gilt ebenfalls für den Effekt der Kategorie Expertenwissen (Modell `logreg-5`). Dieser Effekt ist auf die mittlere positive Korrelation zwischen Daten als Evidenz und Expertenwissen sowie auf die mittlere negative Korrelation zwischen Intuition und Daten als Evidenz zurückzuführen (vgl. Tabelle 12). Da zum einen das Modell `logreg-5` den größten Teil der Variabilität der abhängigen Variablen aufklärt, und zum anderen der Arbeit ein stark theoriegeleitetes Vorgehen zugrundeliegt, sollen die Hypothesen vor dem Hintergrund der multiplen Regressionsmodelle bewertet werden.

In Bezug auf die Richtigkeit der unmittelbar nach dem Experimentieren aufgestellten Hypothese zeigt sich, dass die Hypothesen, die einen Einfluss der Argumentkategorien Intuition (H2.1) und Expertenwissen (H2.2) annehmen, abgelehnt werden müssen. In Bezug auf die Verwendung der Kategorie Messunsicherheiten (explizit) wurde angenommen, dass eine Verwendung dieser Kategorie einen positiven Einfluss auf den Lernerfolg hat (H2.3). Diese Hypothese muss verworfen werden. Stattdessen spricht der beobachtete negative Einfluss auf die Richtigkeit der Hypothese deutlich dafür, dass die Alternativhypothese mit gegenläufigem Effekt zu H2.3 gilt, die daher angenommen wird. Die Verwendung der Kategorie Daten als Evidenz hat entsprechend der Hypothese H2.4 einen positiven Einfluss auf den Lernerfolg und wird daher angenommen.

Die Nachhaltigkeit des Lernerfolgs wurde überprüft, indem in einer Follow-up-Erhebung die Richtigkeit der Hypothese erhoben wurde. Hier zeigt sich, dass die Hypothese in Bezug auf den Einfluss der Argumentkategorie Intuition (H2.5) angenommen werden muss. Die Hypothesen in Bezug auf die Kategorie Expertenwissen (H2.6), Messunsicherheiten (H2.7) und der Kategorie Daten als Evidenz (H2.8) müssen jedoch abgelehnt werden.

5.6 ZUSAMMENFASSUNG DER BEWERTUNG DER STATISTISCHEN HYPOTHESEN

In der nachfolgend aufgeführten Tabelle werden die Bewertungen der statistischen Hypothesen zusammenfassend dargestellt. Wird eine Hypothese angenommen (✓), wird der entsprechende Effektschätzer berichtet. Haben die Ergebnisse gezeigt, dass die Alternativhypothese abgelehnt werden muss, so wird diese Entscheidung entsprechend gekennzeichnet (×). Musste die getestete Alternativhypothese zugunsten der Nullhypothese abgelehnt werden und konnte aufgrund der beobachteten Signifikanz ein zur Forschungshypothese gegenläufiger Effekt beobachtet werden, wird dies durch das Symbol „*“ gekennzeichnet und der entsprechende Effekt berichtet. Es wird in diesen Fällen post hoc die Alternativhypothese angenommen, die den zur ursprünglich formulierten Alternativhypothese entgegengesetzten Zusammenhang bzw. Unterschied annimmt.

Abk.	Hypothese	Bewertung / Effekt
	Statistische Hypothesen bzgl. des Einflusses der personalen Faktoren auf die Stärke der Verwendung der Argumentkategorien (Forschungsfrage 1):	
H1.1	Je niedriger das fachliche Vorwissen, desto mehr argumentieren die Probanden mit der Kategorie Intuition.	✓ ($\beta_{\text{fw}\rightarrow\text{int}} = -0.34$)
H1.2	Je niedriger das fachliche Vorwissen, desto mehr argumentieren die Probanden mit der Kategorie Expertenwissen.	×
H1.3	Je höher das fachliche Vorwissen, desto mehr argumentieren die Probanden mit der Kategorie Messunsicherheiten (explizit).	×
H1.4	Je höher das fachliche Vorwissen, desto mehr argumentieren die Probanden mit der Kategorie Daten als Evidenz.	✓ ($\beta_{\text{fw}\rightarrow\text{evi}} = 0.21$)
H1.5	Je niedriger das Kognitionsbedürfnis, desto mehr argumentieren die Probanden mit der Kategorie Intuition.	×
H1.6	Je niedriger das Kognitionsbedürfnis, desto mehr argumentieren die Probanden mit der Kategorie Expertenwissen.	×
H1.7	Je höher das Kognitionsbedürfnis, desto mehr argumentieren die Probanden mit der Kategorie Messunsicherheiten (explizit).	×

Abk.	Hypothese	Bewertung / Effekt
H1.8	Je höher das Kognitionsbedürfnis, desto mehr argumentieren die Probanden mit der Kategorie Daten als Evidenz.	×
H1.9	Je niedriger die persönliche Relevanz, desto mehr argumentieren die Probanden mit der Kategorie Intuition.	×
H1.10	Je niedriger die persönliche Relevanz, desto mehr argumentieren die Probanden mit der Kategorie Expertenwissen.	$*$ ($\beta_{\text{wdn}\rightarrow\text{exp}} = 0.22$)
H1.11	Je höher die persönliche Relevanz, desto mehr argumentieren die Probanden mit der Kategorie Messunsicherheiten (explizit).	✓ ($\beta_{\text{wdn}\rightarrow\text{mu}} = 0.31$)
H1.12	Je höher die persönliche Relevanz, desto mehr argumentieren die Probanden mit der Kategorie Daten als Evidenz.	×
H1.13	Je niedriger das situationale Interesse, desto mehr argumentieren die Probanden mit der Kategorie Intuition.	×
H1.14	Je niedriger das situationale Interesse, desto mehr argumentieren die Probanden mit der Kategorie Expertenwissen.	×
H1.15	Je höher das situationale Interesse, desto mehr argumentieren die Probanden mit der Kategorie Messunsicherheiten (explizit).	×
H1.16	Je höher das situationale Interesse, desto mehr argumentieren die Probanden mit der Kategorie Daten als Evidenz.	×

Statistische Hypothesen bzgl. des Einflusses Stärke der Verwendung der Argumentkategorien auf den Lernerfolg und dessen Nachhaltigkeit (Forschungsfrage 2):

H2.1	Je mehr die Probanden mit der Argumentkategorie Intuition argumentieren, desto seltener wechseln sie von einer falschen auf eine fachlich adäquate Hypothese.	×
H2.2	Je mehr die Probanden mit der Argumentkategorie Expertenwissen argumentieren, desto seltener wechseln sie von einer falschen auf eine fachlich adäquate Hypothese.	×

Abk.	Hypothese	Bewertung / Effekt
H2.3	Je mehr die Probanden mit der ArgumentKategorie Messunsicherheiten (explizit) argumentieren, desto eher wechseln sie von einer falschen auf eine fachlich adäquate Hypothese.	$*$ ($OR_{\text{mu,post}} = 0.45$)
H2.4	Je mehr die Probanden mit der Argumentkategorie Evidenz argumentieren, desto eher wechseln sie von einer falschen auf eine fachlich adäquate Hypothese.	✓ ($OR_{\text{evi,post}} = 4.31$)
H2.5	Je mehr die Probanden mit der Argumentkategorie Intuition argumentieren, desto niedriger ist die Nachhaltigkeit des Hypothesenwechsels.	✓ ($OR_{\text{int,fu}} = 0.69$)
H2.6	Je mehr die Probanden mit der Argumentkategorie Expertenwissen argumentieren, desto niedriger ist die Nachhaltigkeit des Hypothesenwechsels.	×
H2.7	Je mehr die Probanden mit der ArgumentKategorie Messunsicherheiten (explizit) argumentieren, desto höher ist die Nachhaltigkeit des Hypothesenwechsels.	×
H2.8	Je mehr die Probanden mit der Argumentkategorie Evidenz argumentieren, desto höher ist die Nachhaltigkeit des Hypothesenwechsels.	×

Statistische Hypothesen bzgl. der Unterschiede in der Verwendung der Argumentkategorien zwischen Probanden, die mit dem Computer- bzw. Realexperiment gearbeitet haben (Forschungsfrage 3):

H3.1	Probanden, die mit dem Computerexperiment gearbeitet haben, zeigen bzgl. des Hypothesenwechsels eine stärkere Verwendung der Argumentkategorie Intuition als die Probanden, die mit dem Realexperiment gearbeitet haben.	$*$ ($d_{\text{int,V}} = -0.26$)
H3.2	Probanden, die mit dem Computerexperiment gearbeitet haben, zeigen bzgl. des Hypothesenwechsels eine stärkere Verwendung der Argumentkategorie Expertenwissen als die Probanden, die mit dem Realexperiment gearbeitet haben.	×

Abk.	Hypothese	Bewertung / Effekt
H3.3	Probanden, die mit dem Computerexperiment gearbeitet haben, zeigen bzgl. des Hypothesenwechsels eine geringere Verwendung der ArgumentKategorie Messunsicherheiten (explizit) (explizit) als die Probanden, die mit dem Realexperiment gearbeitet haben.	✓ ($d_{\text{mu,V}} = -1.08$)
H3.4	Probanden, die mit dem Computerexperiment gearbeitet haben, zeigen bzgl. des Hypothesenwechsels eine geringere Verwendung der Argumentkategorie Daten als Evidenz als die Probanden, die mit dem Realexperiment gearbeitet haben.	$*$ ($d_{\text{evi,V}} = 0.18$)

6 DISKUSSION

Zentrales Anliegen der vorliegenden Arbeit ist die Aufklärung eines komplexen Wirkungszusammenhangs zwischen personalen bzw. situationalen Faktoren und der Verwendung bestimmter Argumente beim Aufstellen von Hypothesen auf Grundlage eines Experiments (Forschungsfrage 1 und Forschungsfrage 3) sowie dem daraus resultierenden Lernerfolg in Form der Richtigkeit der Hypothesen und dessen Nachhaltigkeit (Forschungsfrage 2).

Der folgende Abschnitt diskutiert die zentralen Ergebnisse dieser Arbeit vor dem Hintergrund dieser Fragestellungen. Dazu werden zunächst die verschiedenen Teilfragen einzeln diskutiert, um in den letzten Abschnitten das Gesamtergebnis zu betrachten und die zentralen Implikationen der Arbeit abzuleiten. Es werden methodische Limitationen und Grenzen der Interpretation aufgezeigt. Die Diskussion fokussiert dabei insbesondere auf die eigenen Ergebnisse, da es nur selten möglich ist, eine konkrete Verbindung zum Literaturkorpus herzustellen. Zum einen ist der gewählte theoretische Ansatz, die Übertragung des ELM auf das Lernen von Naturwissenschaften durch Experimentieren, bisher einzigartig. Zum anderen wurden in den Vorarbeiten die Argumentationen von Schülern hinsichtlich selbst aufgestellter Hypothesen beim Experimentieren kategorisiert und dort bereits diskutiert. So wurden abhängige Variablen erzeugt, die bisher in anderen Forschungsarbeiten (noch) keine Verwendung fanden.

Es sei angemerkt, dass an dieser Stelle die Testgütekriterien der verwendeten Instrumente nicht diskutiert werden (siehe dazu Abschnitt 4.9 bzw. Anhang B).

6.1 EINFLUSS PERSONALER FAKTOREN AUF DAS ARGUMENTIEREN BEIM EXPERIMENTIEREN

Die Forschungsfrage 1 adressiert den Einfluss verschiedener personaler Faktoren (Fachwissen, Kognitionsbedürfnis, situationales Interesse, Werteinschätzung der Naturwissenschaften) auf die Stärke der Verwendung von vier Argumentkategorien beim Experimentieren (Intuition, Expertenwissen, Messunsicherheiten (explizit), Daten als Evidenz). Zur Operationalisierung dieser Konstrukte wurden standardisierte Verfahren herangezogen, bei denen trotz leichter Modifikationen weiterhin von einer hohen Validität auszugehen ist.

Durch Analyse der latenten Korrelationen zeigte sich zunächst, dass die vier personalen Faktoren überwiegend mittlere bis große Zusammehänge untereinander aufweisen. Aufgrund der großen gemeinsa-

men Varianzanteile ist es daher nicht verwunderlich, dass bei der Untersuchung des Einflusses der personalen Faktoren auf die Verwendung der Argumentkategorien in einem multiplen Regressionsmodell einige der Faktoren keinen signifikanten Beitrag zur Varianzaufklärung in den Argumentkategorien leisten. Dies betrifft zum einen das Kognitionsbedürfnis, welches starke Zusammenhänge mit dem situationalen Interesse, dem Fachwissen Mechanik sowie der Werteinschätzung der Naturwissenschaften aufweist. Zum anderen betrifft dies das situationale Interesse. Hier liegt ein starker Zusammenhang sowohl mit der Werteinschätzung der Naturwissenschaften als auch mit dem Fachwissen Mechanik vor. Sowohl die Effekte des Kognitionsbedürfnisses als auch des situationalen Interesses auf die Stärke der Verwendung der Argumentkategorien erreichen daher keine Signifikanz. Dies widerspricht den Annahmen des ELM, wonach das Kognitionsbedürfnis und das situationale Interesse als motivationale Komponenten einen Einfluss nehmen sollten. Zum einen ist jedoch zu berücksichtigen, dass es sich hierbei um einen methodischen Effekt handelt, da Kognitionsbedürfnis und situationales Interesse von den exogenen Variablen mit der stärksten prädiktiven Kraft aus dem Regressionsmodell „verdrängt" werden. Zum anderen ist insbesondere die Ausprägung der Zusammenhängen zwischen dem situationalen Interesse und den Argumentkategorien ohnehin nur als gering zu bewerten (zwischen .14 und .20, vgl. Tabelle 12). Vergleichbar niedrige Zusammenhänge zwischen motivationalen Komponenten und prozeduralen Größen beim Experimentieren, wie z. B. dem Strategiewissen, berichten Künsting et al. (2008).

Das Fachwissen Mechanik, welches hier als Entsprechung zur Einflussgröße „Fähigkeit zur Verarbeitung einer persuasiven Nachricht" eingesetzt wurde (vgl. Abschnitt 2.4.1), nimmt auf zwei Argumentkategorien Einfluss. Zum einen erklärt das Fachwissen die Verwendung der Kategorie Daten als Evidenz mit einem kleinen positiven Effekt. Zum anderen hat das Fachwissen einen mittleren negativen Einfluss auf die Verwendung der Argumentkategorie Intuition. Beide empirischen Effekte entsprechen dem theoretischen Leitgedanken des ELM: Liegt eine Fähigkeit zur Verarbeitung der Informationen aus einer persuasiven Nachricht vor, wird die Elaborations-Wahrscheinlichkeit positiv beeinflusst, so dass Rezipienten der Nachricht diese eher über die zentrale Route verarbeiten werden. Dies äußert sich hier in einer stärkeren Verwendung der Argumentkategorie Daten als Evidenz. Es ist zu beobachten, dass in der Entscheidung zum Beibehalten bzw. Verwerfen einer zuvor aufgestellten physikalischen Hypothese bei höherem Fachwissen eher Daten als Evidenz herangezogen werden, während gleichzeitig die Argumentkategorie Intuition seltener verwendet wird. Wie in den Vorarbeiten (vgl. Abschnitt 2.4) dargelegt wurde, handelt es sich bei den Kategorien Intuition und Daten als Evidenz um Argumentkategorien, die der peripheren (Intuition) bzw.

zentralen Route (Daten als Evidenz) zugeordnet werden können. Die Ergebnisse dieser Studie unterstützen daher die Annahme, dass – analog zu den Zwei-Prozess-Modellen – (vgl. Abschnitt 2.4) auch beim Lernen von Naturwissenschaften durch Experimentieren zwei Prozesse der Informationsverarbeitung unterschieden werden können: Zum einen eine eher rationale Verarbeitung (hier durch die Verwendung der Kategorie Daten als Evidenz) sowie demgegenüber eine eher nicht-rationale Verarbeitung der Daten (hier durch Rückgriff auf Intuition). Es konnte zudem gezeigt werden, dass zwischen der Verwendung der Kategorien Intuition und Daten als Evidenz ein mittlerer bis starker negativer Zusammenhang vorliegt, was dieser These ebenfalls entspricht.

Weiterhin konnte gezeigt werden, dass die Werteinschätzung der Naturwissenschaften, die hier als Entsprechung zu der im ELM benannten persönlichen Relevanz einer Nachricht eingesetzt wurde, einen mittleren positiven Einfluss auf die Stärke der Verwendung der Kategorie Messunsicherheiten (explizit) hat. Nach dem ELM erhöht die persönliche Relevanz als motivationale Komponente die Elaborationswahrscheinlichkeit. Damit wird eine Verarbeitung von persuasiven Nachrichten über eine zentrale Route bevorzugt. Dieses Ergebnis ist daher konform mit den Annahmen des ELM, da die Verwendung der Argumentkategorie Messunsicherheiten (explizit) einen zentralen Charakter aufweist. Dies wurde in den Vorarbeiten und der Entwicklung des Tests zur Erfassung der Verwendung der Argumentkategorien insbesondere durch die Expertenstudie belegt (vgl. Abschnitt B.5). Aus diesem Ergebnis lässt sich ableiten, dass Lernende beim Experimentieren eher dann Streuung, Variation und Unsicherheiten in Daten erkennen, wenn Sie auf den Skalen zur Einschätzung des Wertes der Naturwissenschaften höher abschneiden. Dieser Effekt ist auch auf Konstruktebene gut zu erklären: Bei Betrachtung der Itemtexte des eingesetzten Instruments (vgl. Abschnitt C.1.4) fällt auf, dass eine Reihe von Items auf die Qualität von Entscheidungen auf Grundlage naturwissenschaftlichen Wissens Bezug nehmen (z. B. „Physik kann mir helfen, in vielen Bereichen meines Lebens bessere Entscheidungen zu treffen“). Auch wenn diese These vor dem Hintergrund der derzeitigen Datenlage noch einer weiteren Überprüfung bedarf, könnte daher angenommen werden, dass eine hohe Werteinschätzung zu einer intensiveren Auseinandersetzung mit Daten und insbesondere den Unsicherheiten führt, weil die Probanden richtige oder „bessere“ Entscheidungen treffen möchten. Neben dem mittleren Effekt auf die Verwendung der Kategorie Messunsicherheiten (explizit) hat die persönliche Werteinschätzung der Naturwissenschaften auch einen kleinen positiven Effekt auf die Verwendung der Argumentkategorie Expertenwissen. Dieser Effekt steht im Widerspruch zu der Annahme, dass die Verwendung der Argumentkategorie Expertenwissen von der Verarbeitung der experimentellen Daten und Beobachtungen

über die periphere Route zeugt. Nach dem ELM wäre dann ein negativer Effekt zu erwarten gewesen (vgl. Hypothese H1.10).

Der beobachtete Effekt lässt sich dann erklären, wenn die Kategorie Expertenwissen nicht der peripheren Route, sondern der zentralen Route der Informationsverarbeitung zugeordnet wird (vgl. Abschnitt 2.4.3). Denn in Lehr-Lern-Situationen ist es durchaus notwendig, sich auf die Expertise von Lehrpersonen zu verlassen, da zum einen nicht alle Sachverhalte eigenständig geprüft werden können und zum anderen Lehrpersonen als Experten ihres Fachs gelten. Dieses Ergebnis stellt zudem einen interessanten Anknüpfungspunkt für weitere Forschungsvorhaben dar. Dabei könnte der Frage nachgegangen werden, ob Lernende nur dann Daten und Beobachtungen aus einem Experiment als Evidenz heranziehen, wenn sie einen gewissen Grad an Expertise erkennen. Darauf deutet auch der relativ große Zusammenhang zwischen den beiden Kategorien hin ($r_{\text{exp}\sim\text{evi}} = .42$). Die Beantwortung dieser Frage wäre hochrelevant, denn so könnten Richtlinien zur Gestaltung von Experimenten abgeleitet werden, die „überzeugend“ wirken, da Lernende einen gewissen Grad an Expertise erkennen und sich in einem argumentativen Prozess entsprechend eher Argumente der Kategorie Daten als Evidenz verwenden, wobei die Verwendung der Kategorie Daten als Evidenz dann wiederum lernförderlich wirken könnte.

Zusammengefasst lässt sich konstatieren, dass sich die Einflüsse von personalen Faktoren auf die Argumentkategorien nur bei drei der vier Kategorien finden lassen. Der Effekt auf die Kategorie Expertenwissen widerspricht den Annahmen des ELMs, ist aber zu erklären, wenn diese Kategorie als zentral betrachtet wird.

Der physikdidaktische Wert dieser Ergebnisse liegt insbesondere in dem empirischen Beleg für den negativen Zusammenhang zwischen Fachwissen und der Kategorie Intuition, sowie dem positiven Zusammenhang zwischen dem Fachwissen und der Kategorie Daten als Evidenz. Im Physikunterricht ist es intendiert, dass Lernende eigene Hypothesen beim Experimentieren auf der Grundlage von experimentellen Daten und Beobachtungen aufstellen, d. h. in einer Argumentation Daten als Evidenz heranziehen. Der positive Zusammenhang zwischen Fachwissen und der Kategorie Daten als Evidenz deutet nun auf eine Möglichkeit hin, wie dies im Unterrichtsgeschehen stärker forciert werden kann. In der vorliegenden Arbeit wurde das Fachwissen Mechanik eingesetzt, um eine Entsprechung zu der aus dem ELM benannten Einflussgröße der Fähigkeit zur Verarbeitung einer Nachricht zu finden (zu den Gründen für diese Entscheidung siehe Abschnitt 2.5). Es sind aber weitere Konstrukte denkbar, die einen ähnlichen oder evtl. sogar noch einen höheren Einfluss hätten. Zum Beispiel wäre es insbesondere beim Einsatz von eher quantitativen Experimenten sinnvoll, dass entsprechende Fähigkeiten zum Umgang mit Daten und Unsicherheiten vor der Durchführung und

Auswertung eines Experiments gefördert werden. Denn wird die Fähigkeit zur Verarbeitung experimenteller Informationen im Vorfeld trainiert, hätte dies einen Einfluss auf die Verwendung von Daten als Evidenz. Gleichzeitig würden Lernende weniger intuitiv argumentieren. Würde diese Interpretation konkret auf das in dieser Studie eingesetzte Experiment zum Fadenpendel bezogen werden, könnte dies bedeuten, dass Schülern vor dem Experimentieren zunächst die Gelegenheit gegeben wird, über Ursachen von Unsicherheiten und über einfache Methoden zur Abschätzung der Unsicherheiten (z. B. Bestimmung der Unsicherheitskomponente durch den Betrag der größten Abweichung vom Mittelwert) zu lernen. Es ist dann davon auszugehen, dass eher Daten als Evidenz herangezogen werden, statt intuitiv für oder gegen eine Hypothese zu argumentieren.

6.2 EINFLUSS SITUATIONALER FAKTOREN AUF DAS ARGUMENTIEREN BEIM EXPERIMENTIEREN

Die Forschungsfrage 3 stellt die Frage, inwieweit sich die Verwendung der Argumentkategorien unterscheidet, wenn Lernende mit unterschiedlichen Lernmedien experimentieren. Zur Beantwortung dieser Teilfrage wurde ein randomisiertes Zweigruppen-Design eingesetzt. Bei der Konstruktion der zwei Lernsettings wurde sichergestellt, dass in beiden Gruppen ein möglichst identischer Experimentierraum zur Verfügung steht, der sich mit Ausnahme von Charakteristika des Mediums (echte Materialien vs. Bildschirmsimulation) auf *inhaltlicher* Ebene nicht unterscheidet (vgl. Abschnitt 4.1.4). Eine für Computersimulationen typische Eigenschaft, nämlich eine beliebig hohe Präzision der Messeinrichtung bei Messwiederholung unter sonst identischen Bedingungen, wurde absichtlich zugelassen. So wurde einem charakteristischen Merkmal von Computerexperimenten Rechnung getragen: Diese sind größtenteils so programmiert, dass sowohl systematische als auch zufällige Messunsicherheiten nicht auftreten. Die Messunsicherheit einer „Messung“, als Parameter zur Charakterisierung der Streuung der Messwerte (Hellwig, 2012), ist in Computersimulationen daher meist null.

Zur Analyse der Mittelwertsunterschiede in der Stärke der Verwendung der Argumentkategorien wurden die beiden Gruppen in MG-SGM verglichen. Es ist dabei hervorzuheben, dass sowohl anhand der Stichprobe der Testentwicklungsstudie (vgl. Anhang B) als auch anhand der Stichprobe der Hauptuntersuchung (vgl. Abschnitt 4.9.7) die Messinvarianz des Tests zur Erfassung der Stärke der Verwendung der Argumentkategorien geprüft wurde. Dies ist hier von besonderer Relevanz. Denn die erhobenen Konstrukte entstehen kurzfristig in situ während der Bearbeitung des Experiments und des verwendeten Fragebogens. Da Messinvarianz vorliegt, können Mittelwerte über die Gruppen hinweg sinnvoll miteinander verglichen werden.

Es wurden Modelle mit und ohne Kontrolle der personalen Faktoren als Kovariaten gerechnet. Die Gruppenvergleiche aus beiden Modellen lassen identische Schlussfolgerungen zu.

Der größte Unterschied beim Arbeiten mit Real- bzw. Computersimulationen besteht in der Verwendung von Argumenten aus der Kategorie Messunsicherheiten (explizit), die mit einem starken Effekt in der Gruppe der Probanden, die mit dem Computerexperiment gearbeitet haben, niedriger ausfällt. Dieser Effekt ist sehr gut zu erklären, denn die beiden Experimente unterscheiden sich inhaltlich lediglich hinsichtlich der Qualität der Daten, d. h. in der Präzision der Messung bei Messwiederholung. Es ist daher anzunehmen, dass der Unterschied in der Verwendung der Argumentkategorie Messunsicherheiten (explizit) auf dieses Merkmal der Computersimulation zurückzuführen ist. Zukünftig bietet dieser Aspekt das Potential für weitere Untersuchungen: Es müsste z. B. geprüft werden, ob sich der Unterschied in der Stärke der Verwendung der Kategorie Messunsicherheiten (explizit) beeinflussen lässt, wenn die „Messung" in dem Computerexperiment weniger präzise Daten ergibt. Es könnte z. B. die verwendete automatische Stoppuhr (vgl. Abschnitt 4.1) durch eine ebenfalls manuell zu bedienende Stoppuhr ersetzt werden. Es ist gut denkbar, dass die Verwendung der Kategorie Messunsicherheiten (explizit) dann weiter absinkt.

Neben dem Unterschied in der Stärke der Verwendung der Kategorie Messunsicherheiten (explizit) konnten signifikante Unterschiede in der Verwendung der Kategorien Intuition und Daten als Evidenz aufgedeckt werden. Diese sind jedoch derart klein, dass eine Interpretation nicht angebracht ist. Vermutlich ist die Signifikanz dieser Mittelwertsunterschiede lediglich auf den großen Stichprobenumfang und die damit verbundene hohe Teststärke zurückzuführen. Dass de facto kein Unterschied in der Verwendung der Kategorie Intuition gefunden werden konnte, widerspricht der Annahme, dass Probanden stärker die Kategorie Intuition verwenden würden, da in Computerexperimenten – im Gegensatz zu Realexperimenten – i. d. R. schneller ein größerer Datensatz aufgenommen werden kann, der dazu führt, dass in der Verarbeitung eher auf Heuristiken zurückgegriffen wird (vgl. Abschnitt 2.6). Dass dieser Effekt nicht empirisch untermauert werden kann, könnte daran liegen, dass die Probanden in beiden Gruppen eine identische Datenmenge aufgenommen haben. Der Experimentierprozess wurde jedoch nicht kontrolliert. Weiterhin sollte die Annahme geprüft werden, ob Probanden, die mit dem Computerexperiment gearbeitet haben, weniger stark die Kategorie Daten als Evidenz heranziehen würden, da es sich – streng genommen – gar nicht um empirische Daten handelt, sondern lediglich um Parameter, die auf der Grundlage eines analytischen Modells generiert wurden. Dieser Unterschied ist ebenfalls derart klein, dass er keine didaktische Relevanz hat. In Bezug auf die Stärke der Verwendung

der Kategorie Expertenwissen wurde angenommen, dass Probanden beim Arbeiten mit einem Computerexperiment aufgrund der bereits in ein Computerprogramm implementierten Expertise eher auf Argumente der Kategorie Expertenwissen zurückgreifen. Auch dieser Effekt lässt sich empirisch nicht belegen. Dies könnte daran liegen, dass schulisches Lernen generell sehr stark von der Expertise einer Lehrperson abhängig ist (Auswahl von Materialien, Büchern, Experimenten). Aus Schülerperspektive ist es durchaus legitim und sinnvoll, sich auf diese Expertise zu verlassen. Dies geschieht in beiden Lernsettings offenbar gleichermaßen.

Auf der Grundlage dieser Ergebnisse lassen sich die folgenden Schlüsse ziehen: Zum einen deutet dieses Ergebnis darauf hin, dass beim Arbeiten mit Computerexperimenten ein wesentlicher Aspekt der naturwissenschaftlichen Erkenntnisgewinnung, nämlich der adäquate Umgang mit der Unsicherheit einer Messung, nicht auftritt bzw. erst gar nicht auftreten kann. Der Umgang mit Messunsicherheiten in Daten ist aber ein Teil von *scientific literacy* (Gott & Duggan, 1996; Gott, Duggan & Roberts, 2014; Heinicke, 2012; Hellwig, 2012). Für den Einsatz von Computerexperimenten in der Schulpraxis bedeutet dies daher, dass Computerexperimente – bedingt durch die Suggestion einer beliebig präzisen Messung – nicht dazu führen, dass Lernende in gleicher Form mit Messunsicherheiten konfrontiert werden, wie dies beim Experimentieren mit echten Materialien der Fall wäre. In der Konsequenz führt dies dann dazu, dass auch im epistemischen Prozess des Argumentierens für oder gegen eine Hypothese dieser Aspekt keine Berücksichtigung findet. Lehrpersonen müssen sich beim Einsatz von Computersimulationen darüber bewusst sein! Nur so kann verhindert werden, dass sich ein unvollständiges Bild oder gar mangelhaftes Bild über die Natur der Naturwissenschaften entwickelt. Die Generalisierbarkeit dieser Aussage muss eingeschränkt werden, denn sie umfasst lediglich quantitative Computerexperimente, die eine beliebig hohe Präzision bei Messwiederholung unter sonst identischen Bedingungen erlauben.

Werden diese Ergebnisse zum einen vor dem Hintergrund der Annahme aus dem ELM, nach der situative Faktoren (d. h. Merkmale der Quelle der persuasiven Information) die Elaborationswahrscheinlichkeit und damit die Verarbeitung über die zentrale bzw. periphere Route beeinflussen, diskutiert, muss konstatiert werden, dass ein Einfluss von Merkmalen, die rein durch das Medium bedingt sind, in dem experimentiert wurde, auf die Stärke der Verwendung der Argumente nicht gezeigt werden konnte. Der Teil des ELM, der einen Einfluss situativer Charakteristika auf die Verarbeitung persuasiver Informationen postuliert, lässt sich daher nicht auf das naturwissenschaftliche Experimentieren übertragen. Der offensichtliche inhaltliche Unterschied zwischen den Lernsettings bzgl. der Datenqualität wirkt sich jedoch aus.

Diese Arbeit hat auch zum Ziel, aufzudecken, inwiefern die Generalisierbarkeit von Befunden zum Lernen durch Experimentieren gegeben ist, die rein auf der Basis von Studien mit Computerexperimenten gewonnen wurden (vgl. Abschnitt 2.2). Es bestand die Hypothese, dass die Art des Lernmediums – real vs. virtuell – die kognitive Verarbeitung von Informationen, die aus Experimenten gewonnen wurden, beeinflusst. Die Ergebnisse dieser Arbeit stützen diese Hypothese jedoch nicht. Abgesehen von dem Unterschied in der Verwendung der Kategorie Messunsicherheiten (explizit) finden sich keine nennenswerten Unterschiede in der Argumentation. Es muss daher weiterhin von der Annahme ausgegangen werden, dass der Einsatz von Computerexperimenten zur Untersuchung verschiedener Fragestellungen zum Experimentieren eine legitime und ökonomische Methode darstellt.

Dieser Befund kann aus physikdidaktischer Perspektive sehr kritisch betrachtet werden: „Messdaten“, die aus Computerexperimenten gewonnen wurden, repräsentieren lediglich Lösungen einer analytischen Modellierung eines physikalischen Phänomens, die durch das zugrundeliegende Computerprogramm unter Berücksichtigung gewählter Parameter reproduziert werden. Die Tatsache, dass keine bedeutsamen Unterschiede in der Stärke der Verwendung der Argumentkategorie Daten als Evidenz auftreten, kann darauf hindeuten, dass Schüler offenbar nicht in der Lage sind, den epistemischen Unterschied zwischen den beiden experimentellen Settings zu erkennen. Aus physikdidaktischer Perspektive wäre aber ein kritischer Umgang mit der Herkunft von Messdaten bzw. der in das Programm implementierten Expertise wünschenswert. Dies spiegelt sich auch in den Bildungsstandards wider, in denen explizit gefordert wird, dass Schüler am Ende der Mittelstufe in der Lage sind, „die Gültigkeit empirischer Ergebnisse und deren Verallgemeinerung“ zu beurteilen (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, 2004, S. 11). Lehrpersonen obliegt daher die Verantwortung, insbesondere beim Einsatz von Computersimulationen, diese Problematik explizit mit Schülern zu thematisieren. Der Einsatz von Computersimulationen kann so auch Gelegenheit bieten, Herkunft und Quelle von Informationen (z. B. aus dem Internet) kritisch zu beurteilen, indem z. B. überprüft wird, welche Eigenschaften der Autor eines Computerexperiments aufweist, oder indem die Ergebnisse eines Computerexperiments mit dem eines anderen verglichen werden. Dies ist oftmals, jedoch nicht immer, relativ leicht möglich (für das hier verwendete Computerexperiment zum Fadenpendel können problemlos eine Reihe von Alternativen gefunden werden).

6.3 EINFLUSS DER ARGUMENTKATEGORIEN AUF DIE RICHTIGKEIT DER HYPOTHESEN

In der vorliegenden Arbeit kann der fachinhaltliche Lernerfolg anhand der Richtigkeit der aufgestellten Hypothese nach dem Experiment beurteilt werden. Von 821 Probanden, die vor dem Experiment eine falsche Hypothese aufgestellt haben, stellen lediglich rund 52 % der Probanden eine fachlich korrekte Hypothese nach dem Experiment auf. Dieses Ergebnis entspricht der Befundlage, nach der der Lernzuwachs an konzeptuell-deklarativem Wissen, der an den Einsatz von Experimenten im naturwissenschaftlichen Unterricht geknüpft ist, häufig hinter den Erwartungen zurückbleibt (Abrahams, 2017; Hofstein & Lunetta, 2004). Die Forschungsfrage 2 adressiert in diesem Zusammenhang den Einfluss der Stärke der Verwendung verschiedener Argumentkategorien auf die Richtigkeit der Hypothese nach dem Experiment. Dieser Ansatz erlaubt die Beantwortung der Frage, inwiefern Prozesse der Verarbeitung von Informationen, die aus einem Experiment gewonnen wurden, das fachliche Lernen beeinflussen. So kann dazu beigetragen werden, aufzudecken, *warum* Lernende Schwierigkeiten haben, konzeptuelles Wissen beim Experimentieren zu konstruieren.

Die fachliche Richtigkeit der Hypothese bzgl. des Zusammenhangs zwischen Schwingungsdauer und Pendelmasse wurde nach dem Experimentieren zu zwei Zeitpunkten erhoben: Unmittelbar nach dem Experimentieren und zum anderen in einer Follow-up-Erhebung 60 bis 90 Tage nach dem Experimentieren. So kann die Annahme des ELM überprüft werden, derzufolge eine Entscheidung, welche überwiegend über die periphere Route der Informationsverarbeitung getroffen wurde, von geringerer Persistenz ist.

Hinsichtlich der Richtigkeit der Hypothese unmittelbar nach dem Experimentieren geht ein großer Effekt von der Stärke der Verwendung der Kategorie Daten als Evidenz aus. Ziehen Lernende beim Experimentieren Daten als Evidenz heran und begründen darauf die Entscheidung zum Aufstellen einer Hypothese, nimmt dies positiven Einfluss auf die Wahrscheinlichkeit bzgl. der Richtigkeit dieser Hypothese. Dieses Ergebnis ist von hoher physikdidaktischer Relevanz, denn zum einen handelt es sich bei der Verwendung dieser Argumentkategorie um einen wünschenswerten, naturwissenschaftlichen und kognitiven Weg der Informationsverarbeitung und des Argumentierens, wie in Abschnitt 2.6 dargelegt wurde. Dieses Ergebnis deutet aber auch darauf hin, dass die Verwendung dieser Kategorie dazu führt, dass der konzeptuelle Lernerfolg durch Experimentieren positiv beeinflusst wird! Gleichzeitig hat aber die Verwendung der Kategorie Messunsicherheiten (explizit) einen mittleren negativen Einfluss auf die Wahrscheinlichkeit bzgl. der Richtigkeit der Hypothese nach dem Experiment. Dieses Ergebnis ist zunächst verwunderlich.

Es hätte erwartet werden können, dass Probanden, indem sie Argumente dieser Kategorie verwenden und somit Messunsicherheiten berücksichtigen, eine fachinhaltlich korrekte Hypothese aufstellen. Dies ist aber offenbar nicht der Fall. Daher ist die Vermutung naheliegend, dass Schüler Unsicherheiten in Daten zwar erkennen, aber aufgrund nicht vorhandener Fähigkeiten zum Umgang mit Unsicherheiten eher bei ihrer falschen Hypothese bleiben. Dieses eher „konservative" Verhalten ist legitim und sinnvoll. Es ist aber zu konstatieren, dass die Konfrontation mit Messunsicherheiten eine kognitive Hürde bei der Argumentation für eine fachlich adäquate Hypothese darstellt! Dieses Ergebnis ist konform mit der Aussage von Kelly (2008), nach der für die erfolgreiche Teilhabe an epistemischen Prozessen in den Naturwissenschaften methodische Fähigkeiten und Kenntnisse über allgemein akzeptierte Kriterien vorliegen müssen („epistemic criteria", S. 103).

In Bezug auf die Nachhaltigkeit des Lernzuwachses konnte gezeigt werden, dass die Verwendung der Argumentkategorie Intuition einen negativen Einfluss auf die Richtigkeit der Hypothese zum Zeitpunkt der Follow-up-Erhebung hat. Verarbeiten Probanden die aus dem Experiment gewonnenen Informationen also eher über die periphere Route, indem sie Argumente aus der Kategorie Intuition verwenden, ist es wahrscheinlicher, dass sie die auf Grundlage des Experiments aufgestellte Hypothese schneller wieder „vergessen".

Vor dem Hintergrund der Annahmen des ELM müssen diese Ergebnisse differenziert diskutiert werden: Der positive bzw. negative Einfluss der Verwendung der Kategorie Daten als Evidenz, welche der zentralen Route der Informationsverarbeitung zugeordnet wurde, bzw. der Kategorie Intuition auf den Lernerfolg entspricht den Annahmen des ELM. Diese Ergebnisse deuten darauf hin, dass es möglich ist, bei der Verarbeitung experimenteller Daten und Beobachtungen – analog zu den benannten Zwei-Prozess-Modellen (vgl. Abschnitt 2.4) – zu differenzieren. Zunächst überraschend ist, dass die Argumentkategorie Messunsicherheiten (explizit) dazu führt, dass Lernende nach dem Experiment eher bei einer falschen Hypothese bleiben. Denn diese Kategorie wurde der zentralen Route zugeordnet, hat aber offenbar einen negativen Einfluss auf die Hypothesenwahl. Vor dem Hintergrund der bereits diskutierten Gründe für diesen Effekt kann dieses Vorgehen aber trotzdem als durchaus rational und damit der zentralen Route der Informationsverarbeitung entsprechend bewertet werden: Denn die Verwendung dieser Kategorie deutet darauf hin, dass Probanden – ganz rational – Unsicherheiten in Daten erkennen, aber dann ihre vor dem Experiment aufgestellte (falsche) Hypothese nicht ablegen. Dieses im Hinblick auf den Umgang mit Hypothesen als konservativ zu bezeichnende Vorgehen ist aber sinnvoll und rational.

Die Verwendung der Kategorie Expertenwissen hat keinen signifikanten Einfluss auf die Wahl der Hypothese und die Nachhaltigkeit der Entscheidung. Dies kann daran liegen, dass hier ein starker Zusammenhang mit Kategorie Daten als Evidenz vorliegt.

Zusammenfassend betrachtet lassen diese Ergebnisse den Schluss zu, dass unterrichtliches Geschehen stärker als bisher Schüler dazu befähigen muss, auf Grundlage selbstständig erhobener Messdaten eigene Hypothesen zu begründen. Aufgrund des negativen Zusammenhangs zwischen Daten als Evidenz und Intuition ist dann zudem anzunehmen, dass sich dies auch auf die Nachhaltigkeit dieser Entscheidung auswirkt. Weiterhin implizieren die Ergebnisse dieser Studie, dass Unterrichtskonzepte und -materialien entwickelt und auf Wirksamkeit überprüft werden müssen, die Schülern ermöglichen, Fähigkeiten bzgl. eines adäquaten Umgangs mit Daten und Messunsicherheiten zu erwerben (vgl. Kelly, 2008; Masnick, Klahr & Knowles, 2017). Denn abgesehen davon, dass es sich dabei um einen grundlegenden epistemischen Prozess in den Naturwissenschaften handelt, ist diese Fähigkeit zwingend notwendig, um die natürliche Variabilität zum einen nicht fälschlich zu interpretieren und zum anderen fachinhaltlich korrekte Hypothesen auf der Basis von experimentellen Daten anzunehmen. Die Ergebnisse von Munier, Merle und Brehelin (2013) deuten darauf hin, dass Lernende adäquate Konzepte zum Umgang mit Daten und Unsicherheit annehmen können.

6.4 LIMITATIONEN

Die Ergebnisse und Interpretationen unterliegen verschiedenen Limitationen. Diese sollen im folgenden Abschnitt kritisch reflektiert werden. Dazu werden Limitationen, die sich aus der verwendeten Methodik ergeben, benannt. Weiterhin wird diskutiert, inwiefern die stark situative und inhaltsspezifische Anlage der Studie die Generalisierbarkeit der Interpretationen einschränkt. Abschließend wird die ökologische Validität bewertet.

6.4.1 *Methodische Limitationen*

Zur Beantwortung der aufgeführten Fragestellungen musste ein Verfahren zur Erfassung der verwendeten Argumentkategorien entwickelt werden. Trotz hinreichend belegter Reliabilität und dem Vorliegen von Evidenzen für verschiedene Aspekte der Validität (vgl. Anhang B), ist es nicht endgültig auszuschließen, dass es dabei zu Verzerrungen kommt. Grenzen des Instruments sind in Anhang B diskutiert.

Ein Ziel der Arbeit war es, den Einfluss der Verwendung verschiedener Argumentkategorien auf die Richtigkeit einer Hypothese zu verschiedenen Messzeitpunkten zu untersuchen. Die Operationalisie-

rung dieser Hypothese wurde realisiert, indem Probanden drei plausible Hypothesen vorgelegt wurden (vgl. Abschnitt 4.3.2). Diesem Verfahren liegen zwei Schwächen zugrunde. Zum einen liegt eine gewisse Wahrscheinlichkeit zum Erraten der richtigen Hypothese vor, welche dieses Verfahren verzerrt. Zum anderen führt dieses Verfahren zu einer binär kodierten Variablen. In der Folge mussten logistische Regressionen gerechnet werden, um den Einfluss der Stärke der Verwendung bestimmter Argumentkategorien auf die Richtigkeit der Hypothese zu bestimmen. Dieses Verfahren führt zum einen zu den eher ungewöhnlich zu interpretierenden Chancenverhältnissen (*odds ratios*). Eine binäre Variable hat zudem einen geringeren Informationsgehalt (es liegt z. B. keine unmittelbar zugängliche Varianz vor). Es ist daher offen, inwiefern sich die hier dargestellten Ergebnisse wiederfinden ließen, wenn statt der Richtigkeit der aufgestellten Hypothese eine Variable „Lernerfolg" auf einem höheren Skalenniveau operationalisiert werden würde, z. B. durch den Einsatz einer (kurzen) Itembatterie, die eher das Verständnis über den untersuchten Zusammenhang abfragt als allein die Richtigkeit des untersuchten Zusammenhangs in binärer Form. Durch ein solches Verfahren könnte diese abhängige Variable differenzierter erhoben werden.

Die aus dem ELM benannte Einflussgröße „Fähigkeit zur Verarbeitung einer Nachricht" wurde durch einen Fachwissentest zum Bereich Mechanik erfasst. Dies geschah insbesondere auf Grundlage einer Reihe empirischer Belege, die darauf hindeuten, dass es sich hierbei um einen in mehrerer Hinsicht bedeutsamen Prädiktor für das Lernen von Physik handelt (vgl. Abschnitt 2.5). Die aufgeklärten Zusammenhänge könnten durch den Einsatz eines Instruments zur Erfassung des Verständnisses über Messprozesse und den Umgang mit Daten in weiteren Studien detaillierter geprüft werden.

Das situationale Interesse wurde in dieser Studie während der Bearbeitung des Fragebogens, nach der Einführung in die Experimentieraufgabe, aber unmittelbar vor der Durchführung des Experiments, erfasst. Es handelt sich daher nicht um das situationale Interesse, das während der Durchführung des Experiments besteht, sondern um das *erwartete* situationale Interesse, das durch die Aufgabenstellung induziert wurde. Entsprechend ist zu prüfen, ob sich der Nulleffekt hinsichtlich des Einflusses des situationalen Interesses auf die Stärke der Verwendung der Argumentkategorien reproduzieren lässt, wenn das situationale Interesse unmittelbar *nach* der Durchführung des Experiments erfasst wird.

Der Vergleich der beiden Experimentalgruppen – real vs. virtuell – wurde sorgfältig geplant, insbesondere wurde bei der Entwicklung der Experimente auf identische experimentelle Möglichkeiten geachtet. Nicht kontrolliert wurde jedoch der experimentelle Arbeitsprozess beider Teilgruppen. Es ist daher in dieser Studie nicht sichergestellt, dass Probanden in beiden Gruppen vergleichbar experimentiert

haben. Hier könnte zukünftig, beispielsweise durch engere Vorgaben bei der Messdatenprotokollierung, sichergestellt werden, dass beide Probandengruppen vergleichbar experimentieren.

6.4.2 *Die inhaltliche und situative Spezifität der Ergebnisse*

Das Argumentieren beim Experimentieren kann als ein epistemisch-kognitiver Prozess gesehen werden, der nicht unabhängig von einer Situation entsteht. Das Einnehmen einer stark situativen Forschungsperspektive auf das Lernen ist in jüngerer Zeit verstärkt gefordert worden, da nur so Lernprozesse und Ergebnisse adäquat untersucht werden können (Chinn et al., 2011; Greeno, 1998; Sandoval, 2012). Für die vorliegende Arbeit bedeutet dies, dass die Ergebnisse stark von den materiellen und inhaltlichen Ressourcen abhängen, mit denen die Probanden während der Durchführung konfrontiert waren – hier die Durchführung eines geeigneten Realexperiments bzw. einer Computersimulation zur Untersuchung des Zusammenhangs zwischen Pendelmasse und Schwingungsdauer bei einem Fadenpendel mit den unter Abschnitt 4.1 beschriebenen Merkmalen. Der Gültigkeitsbereich der Ergebnisse erstreckt sich daher zunächst nur auf diese Situation. Es existieren jedoch Gründe anzunehmen, dass vergleichbare schultypische Experimente, die eine quantitative Aufnahme von Messdaten vergleichbarer Qualität ermöglichen, zu ähnlichen Ergebnissen führen würden (z. B. Bestimmung von g durch Fallversuche, Hookesches Gesetz).

Geprüft werden muss zudem, inwiefern sich die Ergebnisse auf Experimente übertragen lassen, die mit schultypischen Mitteln üblicherweise zu „eindeutigeren“ Ergebnissen führen, da die physikalische Messung interessierender Größen mit geringeren zufälligen Unsicherheiten behaftet ist (wie dies z. B. bei Versuchen zum Ohm'schen Gesetz der Fall wäre). Bei dem hier gewählten Experiment zum Fadenpendel liegt zudem eine Situation vor, in der eine Unabhängigkeit zwischen untersuchten Größen gezeigt werden soll. Für diesen Fall berichten Kanari und Millar (2004) ohnehin eine größere Schwierigkeit als bei Experimenten, die Zusammenhänge zwischen Größen prüfen, die in Wirklichkeit vorliegen.

Obwohl die Studie also eine stark situative Perspektive einnimmt, wurden mit dem Kognitionsbedürfnis und der Werteinschätzung der Naturwissenschaften Konstrukte erhoben, die einen stark kognitiven Charakter haben (Greeno, 1998). Diese Konstrukte sind ohne Bezug auf eine spezifische Situation oder einen bestimmten Inhalt definiert. Dies wird insbesondere bei Betrachtung der vorgenommenen Operationalisierung deutlich. So lautet ein verwendetes Item zur Erfassung des Kognitionsbedürfnisses z. B. : „Die Aufgabe, neue Lösungen für Probleme zu finden, macht mir wirklich Spaß.“ (Item `nfc1` vgl. Abschnitt C.1.2). Dieses Item adressiert keine spezifische Situati-

on, denn es können hier verschiedene „Probleme“ gemeint sein. Es ist plausibel anzunehmen, dass das Kognitionsbedürfnis einer Person von der Situation abhängt, in der sie sich befindet. Diese Eigenschaft der hier verwendeten Operationalisierung des Kognitionsbedürfnisses ist auch von Pechtl (2009) kritisiert worden. Die gleiche Argumentation trifft auch für die hier verwendete Skala zur Erfassung der Werteinschätzung der Naturwissenschaften zu. Zukünftige Forschungsarbeiten könnten daher diese Ausgangslage aufgreifen, indem deutlich stärker auf die Situation bezogene Skalen verwendet werden. Entsprechende Vorschläge liegen mittlerweile für das Kognitionsbedürfnis (Pechtl, 2009) und auch für die persönliche Relevanz bzw. Werteinschätzung vor (z. B. Gaspard, 2015).

Im Zusammenhang mit der stark situativen Spezifität der Ergebnisse muss diskutiert werden, inwiefern die Ergebnisse dadurch verzerrt werden, dass Probanden unter unterschiedlichen epistemischen Zielsetzungen gearbeitet haben können. Chinn et al. (2011) argumentieren, dass unterschiedliche epistemische Zielsetzungen, wie z. B. das Ziel des Verstehens vs. das Ziel des Erreichens eines „minimally justified belief“ (S. 147), dazu führen, dass die prädiktive Kraft von Modellen zur Analyse von Lernprozessen sinkt: Lernende können sich in unterschiedlichen Situationen entsprechend ganz unterschiedlicher Zielsetzungen verhalten. Der Einfluss unterschiedlicher epistemischer Zielsetzungen in dieser Studie kann an einem Beispiel illustriert werden: In der Situation der Datenerhebung ist den Probanden bewusst, dass die Teilnahme an der Studie freiwillig und nicht bedeutend für ihren persönlichen Erfolg im weiteren Verlauf des Physikunterrichts sein wird (auch wenn die Teilnahme an der Studie noch so gut motiviert wurde). Es ist daher zum einen denkbar, dass es Probanden gibt, die z. B. nur ein relativ niedriges Engagement im Hinblick auf die experimentelle Überprüfung der Hypothese zeigen. Probanden unter dieser Zielsetzung könnten eher intuitiv vorgehen. Zum anderen könnte es Probanden geben, die, beispielsweise bedingt durch die Aufgabenstellung, wirkliches Interesse und Bestreben zeigen, durch das Experiment ein Verständnis über den untersuchten Zusammenhang zu erlangen und daher eher die Kategorie Daten als Evidenz verwenden. Anhand dieser beiden Extrembeispiele ist vorstellbar, dass die Stärke der Verwendung der Argumentkategorien in hohem Maße durch unterschiedliche epistemische Zielsetzungen beeinflusst sein könnte. Für zukünftige Arbeiten zum Argumentieren beim Experimentieren könnte dieser Aspekt daher relevant sein, denn bisher liegen kaum empirische Arbeiten vor, die epistemische Zielsetzungen in Untersuchungen zu epistemischen Prozessen mit einbeziehen (Sandoval, 2014).

6.4.3 *Die ökologische Validität der Ergebnisse*

Auch wenn es sich bei dieser Studie nicht um eine Laborstudie im klassischen Sinne handelt, wurde zur Datenerhebung eine vom üblichen schulischen Geschehen abweichende Situation erzeugt: Dies bezieht sich insbesondere auf eine nicht vorhandene Einbettung in den Unterricht durch Vor- und Nachbereitung, auf die Gestaltung der zur Datenerhebung genutzten Unterrichtsstunde durch fremde Personen sowie auf fehlende Merkmale üblichen Unterrichts wie, beispielsweise eine fehlende Zielsetzung der Unterrichtstunde oder eine fehlende Ergebnissicherung. Die Generalisierung der Interpretation der Ergebnisse dieser Studiensituation auf den schulischen naturwissenschaftlichen Regelunterricht muss daher vor diesem Hintergrund bewertet werden. Es ist aber davon auszugehen, dass die Kontextbedingungen, die insbesondere bei der Durchführung des Experiments vorgeherrscht haben, übertragbar sind. Die Schüler führten das Experiment eigenständig während der Bearbeitung des Fragebogens durch, der so konstruiert wurde, dass er weitestgehend eine typische Unterrichtssituation mit Experimenten nachbildet (vgl. Abschnitt C.2). Dies zeichnet sich insbesondere durch die Einbettung in einen Kontext, die Beschreibung der relevanten physikalischen Größen, das Aufstellen einer Hypothese und die Einführung in den Umgang mit den experimentellen Materialien aus. Auch bei der Auswahl des verwendeten Experiments und der Konstruktion der Aufgabe handelt es sich um eine für die Physik der Mittelstufe repräsentative Aufgabe („ecological sampling“ Messick, 1995, S. 745). Es ist daher hier davon auszugehen, dass die ökologische Validität trotz der genannten Einschränkungen hoch ist.

7

FAZIT UND AUSBLICK

„*Lernen, evidenzbasiert zu begründen*“ – Dieses eingangs skizzierte Ziel sollen Lernende durch naturwissenschaftlichen Unterricht erreichen. Die Ergebnisse dieser Arbeit deuten beim Einsatz des Experiments als Argumentationsgelegenheit darauf hin, dass es „Stellschrauben“ gibt, mit denen das evidenzbasierte Argumentieren beim naturwissenschaftlichen Lernen beeinflusst werden kann. Je stärker beispielsweise das fachliche Vorwissen ausgeprägt ist, desto eher ziehen Lernende Daten als Evidenz heran und desto weniger begründen sie intuitiv. In der Folge führt dies wiederum dazu, dass Lernende mit einer höheren Wahrscheinlichkeit fachlich adäquate Hypothesen aufstellen und diese zudem eher länger beibehalten. Basierend auf diesen Ergebnissen erscheint es nötig, Lernumgebungen zu gestalten, in denen das, was naturwissenschaftliche Evidenz ausmacht, nämlich das Heranziehen von relevanten Daten und der Umgang mit Unsicherheiten, explizit erworben werden kann. So könnte das genannte Ziel, das Erlernen evidenzbasierten Argumentierens, erreicht werden.

Der Umgang mit experimentellen Beobachtungen, Messdaten und Unsicherheiten muss explizit gefördert werden!

Diese Aussage birgt eine Reihe von Implikationen für Praxis und Forschung: Es müssen Unterrichtskonzepte und -materialien entwickelt werden, die diese Forderung explizit aufnehmen. Diese könnten in längsschnittlichen, schulüblichen Settings des Physikunterrichts auf ihre Wirksamkeit hin evaluiert werden. Auf diese Weise könnte überprüft werden, inwiefern sich die Aussagen dieser Studie auch unter den Bedingungen im Feld empirisch belegen lassen. Die Entwicklung geeigneter Konzepte und Materialien ist sicherlich eine Möglichkeit, dieser Aussage Rechnung zu tragen. Eine weitere Möglichkeit ist es, die Bedeutung von fachlichem Vorwissen zum Umgang mit experimentellen Daten und Beobachtungen sowie den Einsatz des Experiments als Argumentationsgelegenheit stärker in die Lehrkräftebildung zu integrieren. So können angehende Lehrkräfte lernen, dass die Verarbeitung experimenteller Informationen durch Lernende nicht nur rational, sondern auch intuitiv abläuft und dass die Qualität der Verarbeitung einen Einfluss auf den Lernerfolg durch Experimentieren nimmt.

Eine Reihe von Forschungsarbeiten in der Naturwissenschaftsdidaktik und der Lehr-Lern-Forschung untersucht das Experimentieren durch den Einsatz von Computerexperimenten. Ein weiteres Anliegen dieser Arbeit war es daher zu überprüfen, inwiefern die so gewonnenen Ergebnisse auf das Experimentieren mit echten Mate-

rialien übertragen werden können. Im Rahmen dieser Arbeit kann kein Beleg erbracht werden, dass sich das Argumentieren alleine aufgrund des Lernmediums – real vs. virtuell – bedeutsam unterscheidet. Es gibt jedoch eine Ausnahme: Es konnte gezeigt werden, dass die in Computersimulationen üblicherweise vorgenommenen Vereinfachungen, hier der Verzicht auf Messunsicherheiten, sehr wohl dazu führen können, dass Lernende diesen Aspekt in einer Argumentation für oder gegen eine Hypothese nicht berücksichtigen. Es ist daher davon auszugehen, dass die Übertragbarkeit von Ergebnissen, die in Studien mit Computerexperimenten gewonnen wurden, weiterhin gegeben ist, wenn die verwendeten Computersimulationen inhaltlich – d. h. in Bezug auf die experimentellen Möglichkeiten, die Variablenkontrolle, die Möglichkeiten der Beobachtung, die Datenqualität und die Unsicherheiten – dem Realexperiment entsprechen. Diese Aussage könnte empirisch noch differenzierter überprüft werden. Dazu müsste die hier verwendete Computersimulation hinsichtlich der Qualität der reproduzierten Daten verändert werden. So könnte beispielsweise eine Stoppuhr eingesetzt werden, die manuell mit der Computermaus bedient werden muss. Auf diese Weise würden die gemessenen Werte – unter sonst identischen Bedingungen – bedingt durch die Reaktionszeit die Unsicherheit der Messung erhöhen.

Was bedeutet dieses Ergebnis im Hinblick auf das zweite eingangs skizzierte Ziel naturwissenschaftlichen Unterrichts – „*Lernen über die Natur der Naturwissenschaften*“ – ? Vor diesem Hintergrund besteht die Gefahr, dass der Einsatz von Computersimulationen ohne Messunsicherheiten bei den Lernenden inadäquate Ansichten bzgl. des physikalischen Messprozesses erzeugen kann. Lehrkräfte müssen diesen Aspekt beim Einsatz von Computersimulationen daher thematisieren! Diese Forderung gilt in erweitertem Sinn sicherlich auch für Realexperimente.

Für zukünftige physikdidaktische Forschungsarbeiten gibt es eine Reihe weiterer Anknüpfungsmöglichkeiten, die sich unmittelbar aus der Methodik dieser Studie ergeben. Dies betrifft die Operationalisierung der Hypothesen, die zu verbessernde Situationsspezifität der eingesetzten Instrumente zur Erfassung der personalen Faktoren sowie insbesondere den Zeitpunkt für die Erfassung des situationalen Interesses. Eine weitere Studie mit den vorgeschlagenen Verbesserungen könnte die Ergebnisse aus der vorliegenden Arbeit weiter absichern. Auch würde eine Übertragung der Studie in einen anderen physikalischen Kontext oder eine andere naturwissenschaftliche Domäne wertvolle Erkenntnis darüber liefern, inwiefern sich die hier gezeigten Zusammenhänge über die hier verwendete Situation hinaus generalisieren lassen. Ferner könnte eine Reihe hochinteressanter Fragestellungen untersucht werden, wenn das in dieser Arbeit entwickelte Instrument zu Erfassung der vier Argumentkategorien Intuition, Expertenwissen, Messunsicherheiten (explizit) und Daten als Evi-

denz durch Operationalisierung weiterer Argumentkategorien erweitert werden würde. Hier wäre z. B. insbesondere relevant, inwiefern die eigene Einschätzung bzgl. experimenteller Fähigkeiten (d. h. die Verwendung der Argumentkategorien zur Experimentierkompetenz), das Argumentieren beim Experimentieren beeinflusst.

Das Heranziehen des ELM und die Modellierung des naturwissenschaftlichen Experimentierens als ein Prozess der Verarbeitung persuasiver Informationen ist ein neuer Blickwinkel auf Lernprozesse. Die Frage, inwieweit das ELM auf das Lernen von Naturwissenschaften durch Experimentieren übertragen werden kann, muss differenziert beantwortet werden. Einerseits konnten die Einflüsse situationaler Faktoren empirisch nicht bestätigt werden. Andererseits ließen sich einige Zusammenhänge entsprechend dem ELM empirisch belegen. Dies trifft insbesondere auf die Zusammenhänge zwischen personalen Faktoren und der Verwendung der Argumentkategorien zu. Ferner gelang anhand des ELM der Nachweis, dass die Zentralität bzw. Peripherität einer Argumentation einen theoriekonformen Einfluss auf die Richtigkeit einer Hypothese und Nachhaltigkeit der Entscheidung für diese Hypothese ausübt: Das Argumentieren anhand experimenteller Informationen findet auf einem Kontinuum zwischen rationalen und evidenzbasierten Prozessen auf der einen und einem eher intuitiven Vorgehen auf der anderen Seite statt! Dabei kommt dem rationalen Argumentieren, bei dem Lernende Daten als Evidenz erkennen und verwenden, eine ausgezeichnete Rolle im Lernprozess zu.

Die beschleunigte Expansion des Universums ist auch rund 20 Jahre nach ihrer Entdeckung durch Perlmutter und Kollegen nicht eindeutig erklärt. Die Suche nach der „antreibenden Kraft" führte u. a. zum Konzept der dunklen Energie, die rund 65 % der Masse des Universums ausmachen soll. Worum es sich dabei aber genau handelt, ist ebenso unklar wie die Frage, ob es sie überhaupt gibt (z. B. Copeland, Sami & Tsujikawa, 2006; Josset, Perez & Sudarsky, 2017; Peebles & Ratra, 2003; Scranton et al., 2003; Serra et al., 2009).

Teil II

ANHANG

A

ANHANG VORARBEITEN

Abbildung 8: Versuchsmaterialien zur Bestimmung der Temperatur in einem Festkörper (Abbildung verändert nach Lau, 2013).

Tabelle 22: Beurteilerübereinstimmung und Reliabilität des Verfahrens zur Identifikation der Verwendung verschiedener Argumentkategorien in Schüleraussagen zum Fadenpendel- bzw. Thermometer-Experiment.

	Fadenpendel		Temperatur	
Kategorie	$P_{\text{Ü}}$	ρ[a]	$P_{\text{Ü}}$	ρ
Expertenwissen	96.6 %	1	95.8 %	.59
Daten als Evidenz	82.8 %	.54	47.9 %	.34
Experimentierkompetenz (zentral)	96.6 %	1	97.9 %	.84
Experimentierkompetenz (peripher)	100.0 %	1	97.9 %	[b]
Ignoranz	96.6 %	.85	95.8 %	.75
Intuition	93.1 %	.71	91.7 %	.46
Messunsicherheiten (explizit)	93.1 %	.86	93.8	.88
Messunsicherheiten (implizit)	82.8 %	.71	62.5	.58
Eignung des Experiments	96.6 %	.83	95.8 %	.82
Heranziehen einer (falschen) phys. Theorie	75.9 %	.60	75.0 %	.81

Hinweise: [a] Alle Korrelationen sind signifikant von null verschieden.
[b] Es konnte keine Korrelation berechnet werden, weil ein Beurteiler diese Kategorie nicht beobachtet hat.

B

ENTWICKLUNG EINES TESTS ZUR ERFASSUNG DER STÄRKE DER VERWENDUNG DER ARGUMENTKATEGORIEN INTUITION, EXPERTENWISSEN, MESSUNSICHERHEITEN (EXPLIZIT) UND DATEN ALS EVIDENZ

Dieses Kapitel des Anhangs beschreibt das Vorgehen zur Entwicklung eines fragebogenbasierten Verfahrens zur Erfassung der Stärke der Verwendung der Argumentkategorien Intuition, Expertenwissen, Messunsicherheiten (explizit) und Daten als Evidenz beim Experimentieren. Besonderes Augenmerk wurde dabei zum einen auf die Evaluation der Testgütekriterien gelegt, zum anderen auf ausreichende Testökonomie sowie möglichst hohe Testfairness. Diese Entwicklungsstudie gliedert sich in drei Teile. Zunächst wurde eine Itembatterie entwickelt, die im Folgenden durch ein Expertenrating im Hinblick auf die inhaltliche Validität geprüft wurde. Die reduzierte Itembatterie wurde an einer Stichprobe der für die Hauptuntersuchung relevanten Population empirisch evaluiert.

B.1 ENTWICKLUNGSZIELE DER STUDIE

Das übergeordnete Ziel der Studie ist die Entwicklung eines validen und ökonomischen Instruments zur reliablen Erfassung der vier ausgewählten Argumentkategorien auf der Basis eines Likert-skalierten Paper-Pencil-Verfahrens. Die Studie verfolgte daher die folgenden Entwicklungsziele:

ZIEL 1: ENTWICKLUNG EINER ITEMBATTERIE Auf der Basis der codierten Interviewdaten aus den Vorarbeiten (Abschnitt 2.3) soll eine ausreichende Zahl von Items je Argumentkategorie entwickelt werden. Bei der Entwicklung soll ergänzend auf bereits bestehende Skalen zurückgegriffen werden. Diese Möglichkeit besteht insbesondere bei der Argumentkategorie Intuition (z. B. Betsch, 2004; Keller et al., 2000). Dieses Vorgehen ist nach Bühner (2010) als rationale bzw. deduktive Methode der Testkonstruktion zu bezeichnen.

ZIEL 2: EVALUATION DER INHALTLICHEN VALIDITÄT Obwohl die neu zu entwickelnden Items auf der Basis von bereits codierten Interviewdaten abgeleitet werden sollen, ist erneut zu prüfen, ob die Aussagen, wie beabsichtigt den entsprechenden Argumentkategorien inhaltlich zuzuordnen sind.

ZIEL 3: EVALUATION DER PSYCHOMETRISCHEN QUALITÄT Das inhaltlich validierte Set an Items soll zunächst anhand von Kriterien der klassischen Testtheorie wie Itemschwierigkeit, Itemvarianz, Trennschärfe und Selektionskennwert (z. B. Lienert & Raatz, 1998) auf die psychometrische Qualität hin überprüft und weiter verkleinert werden. Weiterhin sollen Hinweise für das Vorliegen von Konstrukt- bzw. faktorieller Validität gesammelt werden. Zur Evaluation der Konstruktvalidität eines psychometrischen Tests werden im Allgemeinen theoretische Annahmen mittels geeigneter statistischer Methoden anhand empirischer Daten überprüft (Hartig et al., 2012). In der vorliegenden Arbeit sollen vier der zehn Argumentkategorien weiter operationalisiert und in Folgestudien näher untersucht werden (siehe Abschnitt 2.5.3). Es ist daher theoretisch eine vierfaktorielle Struktur des Testentwurfs zu erwarten.

Ferner soll geprüft werden, ob das Set von Items die Voraussetzungen für sinnvolle Vergleiche von Mittelwerten aus verschiedenen Populationen erfüllt (hier der Vergleich der Gruppe der Probanden die mit dem realen Experiment gearbeitet haben mit der Gruppe der Probanden die mit dem virtuellen Experiment gearbeitet haben). Eine Begründung zu diesem Vorgehen findet sich im Methodenteil des Haupttextes Abschnitt 4.3.1.Durch Prüfung der Messinvarianzbedingungen wird im Wesentlichen den folgenden Fragestellungen nachgegangen (Brown, 2006; Vandenberg & Lance, 2000):

A. Werden in beiden Gruppen identische Konstrukte erfasst (d. h. liegt gleiche Faktorstruktur vor)?

B. Sind die Zusammenhänge zwischen der Ausprägung der latenten Variablen und dem Antwortverhalten auf manifester Ebene und damit die Messbeziehungen gruppenübergreifend identisch (d. h. liegen identische Faktorladungen vor)?

Aus einer fehlenden Messinvarianz können verzerrte Schätzer für den wahren Wert des zu schätzenden Parameters resultieren. Sollte sich zeigen, dass die Messbeziehungen zwischen Items und latenten Faktoren über die Gruppen hinweg nicht identisch sind, können Indikatoren identifiziert werden, die einem starken Bias unterliegen, d. h. unterschiedliches Antwortverhalten bei gleichem wahren Wert auf dem latenten Faktor in unterschiedlichen Gruppen hervorrufen. Diese Items können im Folgenden aus weiteren Analysen ausgeschlossen werden, um Messinvarianz herzustellen. Nur beim Vorliegen von Messinvarianz können Mittelwerts- und Strukturunterschiede in Folgestudien sinnvoll interpretiert werden (Brown, 2006; Vandenberg & Lance, 2000). Im Kontext der vorliegenden Arbeit wird hier geprüft, ob der Faktor, der die Ausprägung einer Argumentkategorie in der Gruppe der mit dem realen Aufbau experimentierenden Probanden repräsentiert, vergleichbar ist mit dem Faktor, der die Ausprägung

der Argumentkategorie in der Simulationsexperiment-Gruppe repräsentiert.

B.2 ARBEITSDEFINITIONEN DER ZU ERFASSENDEN MERKMALE

Auf der Basis der Kategorienbeschreibungen aus den Vorarbeiten (Abschnitt 2.3) wurde zunächst eine präzisere Arbeitsdefinition bzw. Konstruktbeschreibung für jede der vier ausgewählten Argumentkategorien erstellt (Bühner, 2010). Diese Konstruktbeschreibung soll zum einen eine präzise Itemformulierung erlauben, sowie als Norm für eine nachfolgende inhaltliche Validierung dienen. Die Arbeitsdefinitionen sind im Folgenden dargestellt.

INTUITION Die Kategorie *Intuition* umfasst alle Aussagen, die darauf schließen lassen, dass die Entscheidung zum Wechseln bzw. Beibehalten einer Hypothese ohne diskursiven Gebrauch des Verstandes, sondern durch eher gefühlsmäßige Eingebungen und Ahnungen zustande kommt. Die Entscheidung wird ohne expliziten Rückgriff auf die experimentellen Beobachtungen bzw. Daten getroffen. Argumentativ berufen sich Aussagen dieser Kategorie oft auf das ‚Bauchgefühl'. Auch ‚unbewusste' und nicht benannte bzw. benennbare Aspekte führen zu einer intuitiven Argumentation.

EXPERTENWISSEN Die Kategorie *Expertenwissen* umfasst Aussagen, die das Wissen eines ausgewiesenen Experten (im Experiment) berücksichtigen. Im Falle der Experiment-Simulation meint dies eine Bezugnahme auf das zugrundeliegende analytische Modell der physikalischen Realität und berücksichtigt die Tatsache, dass dieses bereits durch eine Person mit größerem physikalischen Verständnis programmiert wurde. Im Falle des Realexperiments umfasst diese Kategorie Aussagen, die z. B. das Vertrauen in den experimentellen Aufbau in Bezug auf die Herkunft des Experiments (z. B. Wissenschaftler an einer Universität) bewerten.

MESSUNSICHERHEITEN (EXPLIZIT) Diese Kategorie umfasst Aussagen, die darauf schließen lassen, dass eine Auseinandersetzung im Hinblick auf die Verlässlichkeit der Interpretation der Messdaten stattfindet. Diese kritische Herangehensweise kann sich auf „schwankende" Messwerte bei der Wiederholung des Experiments oder auf die Genauigkeit des Messprozesses beziehen. Diese Kategorie beinhaltet keine Aussagen, die sich aus einer Beurteilung der eigenen experimentellen Fähigkeiten ergeben und sich damit auf personenbezogene Unsicherheitsquellen beziehen.

DATEN ALS EVIDENZ Die Argumentkategorie *Daten als Evidenz* umfasst alle Aussagen, die darauf schließen lassen, dass die Entscheidung zum Beibehalten oder Wechseln der Hypothese auf Evidenzen beruhen. Im Allgemeinen werden Evidenzen als empirische Befunde verstanden, die (wissenschaftliche) Erkenntnis rechtfertigen (vgl. Abschnitt 2.3). Im vorliegenden Kontext des selbstständigen Experimentierens wird daher z. B. konkret auf experimentell ermittelte Daten bzw. Messwerte Bezug genommen. Es ist zu beachten, dass diese Kategorie sowohl Aussagen umfasst, die konkret auf gemessene Daten verweisen, als auch Aussagen, welche die gemessenen Daten nur implizit berücksichtigen („Es wurde ja durch das Experiment bewiesen, dass die Zeit immer gleich war“).

Weiterhin wurde eine detailliertere Beschreibung der vorgenommenen Klassifizierung hinsichtlich der Zentralität und Peripherität der Argumentkategorien angefertigt.

ZENTRALE KLASSE Die zentrale Klasse umfasst diejenigen Argumentkategorien, die auf eine rationale, d. h. vernunftgeleitete und sachliche Auseinandersetzung mit dem gesamten experimentellen Prozess (Durchführung des Experiments, Beobachtung und Sammeln der Daten sowie Evaluation der Daten) schließen lassen. Die zentrale Klasse zeichnet sich besonders durch eine kritische, intensive, aktive und motivierte Auseinandersetzung mit den Informationen aus. Argumentationen aus dieser Klasse stützen sich besonders auf Beobachtungen, Fakten und Evidenzen und sind daher meist intersubjektiv verständlich, d. h. für andere logisch nachvollziehbar und überprüfbar.

PERIPHERE KLASSE Die periphere Klasse umfasst hingegen alle Kategorien, die auf eine eher nichtrationale Auseinandersetzung mit dem experimentellen Prozess schließen lassen. Argumentationen in dieser Klasse orientieren sich oftmals an sog. peripheren Hinweisreizen (Cues), indem sie das Vertrauen in die Informationen und die Herkunft der Informationen analysieren und sich darauf beziehen. Argumentationen dieser Klasse können ferner auf einer gefühlsgeleiteten, intuitiven Auseinandersetzung mit dem experimentellen Prozess beruhen und sind allgemein geprägt von unkritischer und einer eher wenig rationalen Auseinandersetzung mit den Informationen.

B.3 ITEMENTWICKLUNG

Aus der Festlegung der Zielgruppe auf Schüler der Jahrgangsstufe 8 und 9 lassen sich zudem Kriterien der Itementwicklung ableiten. So ist insbesondere auf sprachlicher Ebene auf Schülerverständlichkeit

zu achten. Z. B. sollen Wörter wie „Hypothese“ durch „Vermutung“ ersetzt werden.

Neben der Berücksichtigung der Sprachbeherrschung der Zielgruppe (Bühner, 2010), der Vemeidung von allgemeinen Präferenzen („bei den meisten Entscheidungen berücksichtige ich...“) wurde bei der Entwicklung der Items möglichst darauf verzichtet, negativ gepolte Items zu formulieren, da die Itemtexte eher lang sind, und damit die Gefahr des „Überlesens“ der Invertierung besteht. Vorteile und Probleme negativ formulierter Items werden in der Literatur kontrovers diskutiert (Weijters & Baumgartner, 2012). Insgesamt wurden lediglich fünf invertierte Items formuliert.

Zur Entwicklung der Items wurde neben den Kategorienbeschreibungen das umfassende Interviewmaterial aus den Vorarbeiten herangezogen (vgl. Abschnitt 2.3). Dabei wurden aus den kodierten Schüleraussagen Itemtexte so umformuliert, dass sie den o. g. Ansprüchen genügen, also sowohl beim Beibehalten als auch Verwerfen einer Hypothese sinnvoll beantwortet werden können. Eine exemplarische Gegenüberstellung von Interviewsegment und Itemtext stellt Tabelle 23 dar. Neben Validitätsaspekten (siehe Abschnitt B.5) kann so außerdem die sprachliche Passung sichergestellt werden, da Itemtexte aus Schüleraussagen generiert wurden.

Des Weiteren wurde bei der Entwicklung von Items für die Skalen Intuition und Daten als Evidenz auf Subskalen bereits entwickelter und dokumentierter Instrumente zur Erfassung rationaler und intuitiver Entscheidungs- bzw. Persuasionsprozesse wie der deutschsprachigen Fassung des *Rational-Experiential Inventory* (Epstein et al., 1996; Keller et al., 2000) zurückgegriffen. Ergänzend dazu wird das ebenfalls deutschsprachige Inventar *Präferenz für Intuition und Deliberation* (PID) herangezogen (Betsch, 2004). Bei allen genannten Instrumenten ist zu beachten, dass in der vorliegenden Form kognitive und behaviorale Einstellungen erfasst werden. Die Testitems können also lediglich als Grundlage zur Umformulierung dienen. Das folgende Beispiel zeigt, wie die Itementwicklung auf Basis der etablierten Inventare vorgenommen wurde. Das Item

Bei den meisten Entscheidungen ist es sinnvoll, sich ganz auf sein Gefühl zu verlassen.
(Item #4 entnommen aus Betsch, 2004, S. 183)

wurde umformuliert zu

Bei der Entscheidung zum Beibehalten/Verwerfen meiner Vermutung verlasse ich mich auf mein Gespür. (Argumentkategorie Intuition, Item `i1.11`)

Auch bei der Entwicklung von Items zur Argumentkategorie Daten als Evidenz konnten Items aus etablierten Instrumenten umformuliert werden. So wurde hier z. B. das Item

Bevor ich Entscheidungen treffe, denke ich meistens erst mal gründlich nach.
(Item #14 entnommen aus Betsch, 2004, S. 183)

umformuliert zu

Ich denke erst über die Messdaten nach, bevor ich eine Entscheidung treffe.
(Argumentkategorie Daten als Evidenz, Item `i8.12`)

Tabelle 23: Exemplarische Gegenüberstellung von Interviewsegmenten und daraus abgeleiteten Items.

PROBAND	AUSSAGE	ABGELEITETES ITEM
Expertenwissen		
176AS03	„Weil das Experiment, was sie uns gegeben haben wird ja nicht falsch sein. Das Experiment wird ja nicht falsch sein, also dass das irgendwie falsche Ergebnisse sagt.“	Bei meiner Entscheidung berücksichtige ich, dass das Experiment ja bestimmt nicht falsch sein wird. (i6.14)
155SO05	„Weil ich euch vertraue, wenn es nicht darauf hinausläuft dass ich mich irgendwie oder der Computer mich nicht täuscht.“	Ich vertraue auf den Entwickler des Experiments. Das beeinflusst meine Entscheidung, die Vermutung zu verwerfen / beizubehalten. (i6.16)
Messunsicherheiten (zentral)		
145KA03	„Nicht ganz genau, aber lag schon immer nah beieinander. Diese kleinen Abstände würden mich davon abhalten zu sagen ‚sehr sicher‘.“	Bei meiner Entscheidung zum Beibehalten / Verwerfen meiner Vermutung berücksichtige ich, dass die Messwerte deutlich schwanken. (i7b.10)
Daten als Evidenz		
157SA02	„Weil wir das gerade gemessen haben. “	Die Auswertung meiner Messung ist der Grund
179AN07	„Ich habe gemerkt, dass die Zeit immer gleich geblieben ist, egal wie groß die Masse ist. Ich vermute das, weil ich das so gemessen habe.“	für meine Entscheidung zum Beibehalten / Wechseln meiner Vermutung. (i8.18)
165HE04	„Weil ich das prüfen konnte. Weil es im Experiment so war, kann aber auch sein, dass es im Experiment so war. “	Anhand der Messwerte überprüfe ich, ob ich meine Vermutung wechsle oder beibehalte. (i8.22)

Nach dem hier geschilderten Vorgehen konnte eine Batterie von insgesamt 88 Items entwickelt werden (25 für die Kategorie Intuition, 16 Expertenwissen, 21 Messunsicherheiten (explizit), 26 Daten als Evidenz), die Eingang in die inhaltliche Validierung fanden.

B.4 VORÜBERLEGUNGEN ZUR ANALYSE DER VALIDITÄT DES TESTENTWURFS

Der folgende Abschnitt beschreibt in Bezug auf die Entwicklungsziele 2 und 3 das Vorgehen zur Evaluation der Validität und der psychometrischen Qualität des Testentwurfs (vgl. Abschnitt B.1).

Vereinfacht ausgedrückt handelt es sich bei der Validität eines Tests um die Eigenschaft des Tests, das zu messen, was er zu messen vorgibt (Bühner, 2010). Während traditionelle Konzeptionen eine Unterteilung des Validitätsbegriffs in Inhalts-, Kriteriums- und Konstruktvalidität vornehmen (wie z. B. in Bühner, 2010; Schmiemann & Lücken, 2013) sehen modernere Beiträge Validität als ein „integriertes bewertendes Urteil über das Ausmaß, in dem die Angemessenheit und die Güte von Interpretationen und Maßnahmen auf Basis von Testwerten [...] durch empirische Belege und theoretische Argumente gestützt sind“ (Messick, 1989, S. 13, zitiert nach Hartig et al., 2012, S. 144). In diesem Verständnis sollen im folgenden Abschnitt auf der Basis theoretischer Überlegungen und empirischer Evidenzen Argumente für das Vorliegen von Validität dargelegt werden.

Da die in Abschnitt 4.3.1 festgelegte Methodik zur Erfassung der Verwendung der Argumente, nämlich die fragebogenbasierte Selbstauskunft anhand von Likert-skalierten Itemstimuli post hoc nach dem Experimentieren zunächst sehr ungewöhnlich erscheint, wird in der vorliegenden Arbeit ein besonderes Augenmerk auf eine ausführliche Evaluation verschiedener Aspekte der Validität gelegt.

B.5 EVALUATION DER INHALTLICHEN VALIDITÄT DES TESTENTWURFS

Die Inhaltsvalidität beschreibt, inwiefern Item*inhalte* tatsächlich das zu messende Zielkonstrukt erfassen (z. B. Bühner, 2010; Hartig et al., 2012). Bei den in der vorliegenden Arbeit interessierenden Zielkonstrukten, wie bspw. der Verwendung bestimmter Typen von Argumenten beim Beibehalten bzw. Verwerfen einer physikalischen Hypothese vor dem Hintergrund eines selbstständig durchgeführten physikalischen Experiments, handelt es sich um theoretisch definierte Merkmale (Hartig et al., 2012, S. 150, siehe auch Abschnitt 4.3.1). Dabei wird aus dem Antwortverhalten auf das zugrunde liegende, latente Merkmal geschlossen. Es gilt die Annahme, dass Unterschiede im Antwortverhalten durch Unterschiede im latenten Merkmal erklärt werden können. Die Rechtfertigung dieses Schlusses steht im Zen-

trum der Evaluation der Inhaltsvalidität bei theoretisch definierten Merkmalen. Auf Itemebene muss daher geprüft werden, ob das zu messende Konstrukt Unterschiede im manifesten Antwortverhalten (hier das Zustimmen bzw. Ablehnen einer Aussage) erklären kann, d. h. hier muss geprüft werden, ob eine inhaltliche Passung der Itemtexte im Sinne der bei der Entwicklung intendierten Argumentkategorie vorliegt (Hartig et al., 2012).

Durch theoretische Überlegungen kann zunächst folgendes Argument für das Vorliegen von inhaltlicher Validität des Testentwurfs entwickelt werden: Die Itembatterie wurde auf Basis des Interviewmaterials aus den Vorarbeiten (Abschnitt 2.3) entwickelt. Die Interviews wiederum können als verbalisierte Argumentationen hinsichtlich des intern ablaufenden Entscheidungsprozesses zum Wechseln oder Beibehalten einer Hypothese beim selbstständig-experimentellen Generieren (widersprüchlicher) Daten aufgefasst werden. Die Interviews beinhalten daher bereits manifestes Verhalten der latenten Konstrukte „Verwendung bestimmter Argumente". Die Entwicklung von Itemtexten aus den vorliegenden Interviews stellt daher bereits ein Argument für hohe inhaltliche Validität dar, insbesondere weil die Schüleraussagen bereits mit einer hohen Reliabilität kodiert, d. h. entsprechenden Argumentkategorien zugeordnet werden konnten.

Die zwangsläufig notwendige Umformulierung der Schüleraussagen zu Itemtexten macht jedoch eine erneute Überprüfung notwendig. Um die inhaltliche Passung der Items detaillierter zu evaluieren, wurde daher ein Expertenrating durchgeführt (Hartig et al., 2012; Schmiemann & Lücken, 2013). Übergeordnete Fragestellung war dabei, ob die Items sowohl im Hinblick auf die Zugehörigkeit zur entsprechenden Argumentkategorie als auch im Hinblick auf die Zentralität und Peripherität von Experten entsprechend eingeordnet werden können.

B.5.1 *Methode*

Als geeignete Experten wurden fortgeschrittene Lehramtsstudenten in einem naturwissenschaftlichen Fach bzw. wissenschaftliche Mitarbeiter in einer jeweiligen Fachdidaktik definiert. Es wurde eine Gelegenheitsstichprobe mit $n = 8$ gezogen (Alter 26 bis 34 Jahre). Der akademische Hintergrund der Teilnehmer ist dabei wie folgt zu beschreiben: Drei der Experten haben Physik und Mathematik auf Lehramt studiert und waren zum Zeitpunkt der Untersuchung kurz vor Abschluss ihres Studiums, je ein Teilnehmer hat die Kombination Chemie und Mathematik bzw. Chemie und Geographie studiert (beides wissenschaftliche Mitarbeiter in einer Naturwissenschaftsdidaktik), ein weiterer Teilnehmer Englisch und Biologie. Ein Teilnehmer hat ein Lehramtsstudium für Grundschulen absolviert, ein weiterer ein geisteswissenschaftliches Studium abgeschlossen. Beide arbeiten

als wissenschaftliche Mitarbeiter in einer Naturwissenschaftsdidaktik.

Durch das Heranziehen von Experten aus verschiedenen naturwissenschaftlichen Domänen soll die inhaltliche Gültigkeit der Itembatterie über das Experimentieren in der Physik hinaus sichergestellt werden.

Die Befragung wurde fragebogenbasiert durchgeführt, der Fragebogen ist in Abschnitt B.8.1 dargestellt. Im ersten Teil des Fragebogens wurden die Experten in das Forschungsprojekt eingeführt und mit der Definition der zentralen und peripheren Klasse der Argumente vertraut gemacht. Auf der Basis dieser Definition wurden die Probanden zunächst aufgefordert die Items hinsichtlich ihres Verarbeitungscharakters in die zentrale bzw. periphere Klasse einzuordnen. Im weiteren Verlauf des Fragebogens wurden die Experten über die Argumentkategorien aufgeklärt und ordneten die Items dann je einer der vier Kategorien zu. Durch dieses zweischrittige Verfahren ist sichergestellt, dass die Experten ihre Urteile hinsichtlich der Peripherität bzw. Zentralität sowie hinsichtlich der zugrundeliegenden vier Argumentkategorien unabhängig fällen. In beiden Testteilen war außerdem auch die Antwortmöglichkeit „keine Antwort trifft zu“ zugelassen.

Um einen Hinweis auf die Qualität der Methode zu erhalten, wurde die Itembatterie um vier Items ergänzt, die nicht eindeutig einer Kategorie zugeordnet werden konnten (z. B. „Ich habe mich bei dieser Entscheidung auf mein Bauchgefühl und die experimentellen Beobachtungen gestützt.“). Es ist zu erwarten, dass ein solches mehrdeutiges Item nicht zuzuordnen ist.

Die Bearbeitungszeit für den gesamten Fragebogen betrug 90 Minuten, die jedoch kein Proband vollständig ausschöpfte.

B.5.2 *Datenaufbereitung und -analyse*

Die Fragebögen wurden in eine tabellarische Form überführt. Dabei wurden die Antworten numerisch codiert. Es liegt nominales Skalenniveau vor.

B.5.3 *Ergebnisse*

Zur Auswertung der Expertenbeurteilungen wurde zunächst ein Kriterium definiert, um mangelnde inhaltliche Passung einzelner Items zu identifizieren. Es wurde festgelegt, dass inhaltliche Validität eines Items dann gegeben ist, wenn in beiden Teilen des Tests sieben der acht Experten das Item übereinstimmend bewerten. Dabei wurde auch geprüft, dass diese Übereinstimmung theoriekonform ist, die Experten also nicht nur zufällig in einer anderen Kategorie übereinstimmen. Aus diesem Kriterium lässt sich die paarweise prozentuale

Übereinstimmung $P_{\text{paarweise}}$ je Item und Testteil berechnen. Diese ist definiert als die Häufigkeit, mit der zwei Experten identisch urteilen, dividiert durch die Anzahl der Expertenkombinationen bei n Experten $\frac{n(n-1)}{2}$ und die Gesamtzahl der bewerteten Objekte, die hier eins beträgt, da die $P_{\text{paarweise}}$ auf Itemebene bestimmt wird (Wirtz & Caspar, 2002, S. 47). Die paarweise prozentuale Übereinstimmung bei sieben von acht identischen Beurteilungen ergibt sich daher hier zu

$$P_{\text{paarweise, 7 von 8 identisch}} = \frac{7 \cdot 6/2}{8 \cdot 7/2} \cdot 100\,\% = 75\,\%. \tag{22}$$

Bei der Berechnung der Beurteilerübereinstimmung ist es im Allgmeinen nicht ausreichend, lediglich Maße anzugeben, die nicht um die zufällige Übereinstimmung korrigiert sind (z. B. Wirtz & Caspar, 2002). Etablierte Übereinstimmungskoeffizienten wie Fleiss' κ berücksichtigen daher die Wahrscheinlichkeit, mit der Beurteiler auch zufällig zu einer gleichen Einschätzung gelangen können. In der vorliegenden Arbeit ist die zufällige Wahrscheinlichkeit aufgrund der hohen Zahl der Beurteiler jedoch sehr gering: Im ersten Testteil stehen den Experten drei Antwortmöglichkeiten zur Auswahl („peripheres Item", „zentrales Item", „keine Antwort trifft zu"). Daher ist die Wahrscheinlichkeit, dass sieben von acht Experten zufällig übereinstimmen, gegeben zu

$$P_{\text{Zufall}} = \left(\frac{1}{3}\right)^7 \cdot 3 \approx 0.0014 \tag{23}$$

und damit vernachlässigbar klein. Im zweiten Testteil sinkt die zufällige Übereinstimmungswahrscheinlichkeit sogar noch weiter ab, da fünf Antwortmöglichkeiten vorliegen: $P_{\text{Zufall}} = \left(\frac{1}{5}\right)^7 \cdot 5 \approx 6.4 \cdot 10^{-5}$.

Das nachfolgende Beispiel verdeutlicht noch einmal das Verfahren: Ein Item wurde nur dann eingeschlossen, wenn es hinsichtlich der Zentralität bzw. Peripherität von sieben von acht Experten eindeutig bewertet, sowie hinsichtlich der Zuordnung zu einer Argumentkategorie ebenfalls von mindestens sieben von acht Experten richtig zugeordnet wurde.

Im Laufe der Analyse wurde dieses Kriterium gesenkt, da insbesondere für die Skala Expertenwissen zu wenige Items dieses Kriterium erreichten. In Einzelfällen wurden dann Items selektiert, die in einem der beiden Testteile eine paarweise prozentuale Übereinstimmung $P_{\text{paarweise}} = 75\,\%$, im jeweils anderen Teil aber lediglich einen Wert von $P_{\text{paarweise}} = 71.4\,\%$ erreichten, was einem identischen Urteil von sechs von acht Experten entspricht. Dieses aufgeweichte Kriterium wurde nur bei sechs Items angewendet.

Für die Skala Intuition konnten zunächst 22 inhaltlich valide Items identifiziert werden. Die Itembatterie wurde für den weiteren Verlauf der Testentwicklung um die Items `i1.13` und `i1.14` reduziert, da diese äußerst ähnlich zu anderen Items (z. B. `i1.31`) sind. Ferner wurden

die Items i1.22 sowie i1.23 ausgeschlossen, da sie zwar inhaltlich der Argumentkategorie Intuition zuzuordnen sind, aber die Formulierung der Items zudem nach dem Vertrauen in die Daten bzw. in das Experiment fragt. Dieser Aspekt ist definitionsgemäß durch die Argumentkategorie Intuition nicht abgedeckt. In die weitere Testentwicklung gehen für die Skala Intuition daher 18 Items ein.

Für die Skala Messunsicherheiten (explizit) konnten 12 Items mit $P_{\text{paarweise}} = 75\,\%$ selektiert werden. Das Item i7b.06 wurde von allen Experten in die zentrale Klasse eingeordnet, jedoch gaben zwei Experten an, dass das Item keiner Kategorie zugeordnet werden könne. Die Zahl der Items in der Skala Messunsicherheiten zentral liegt daher bei 13.

Für die Skala Daten als Evidenz konnten zunächst 17 inhaltlich valide Items identifiziert werden. Drei korrekt klassifizierte Items wurden lediglich von sechs Experten in die korrekte Argumentkategorie eingeordnet. Der Itempool für die Skala Daten als Evidenz wurde für den weiteren Verlauf der Testentwicklung um diese drei Items erweitert.

Für die Skala Expertenwissen erfüllten nur vier Items das Kriterium $P_{\text{paarweise}} = 75\,\%$, zwei Items (i6.08 und i6.13) erfüllten das abgeschwächte Kriterium $P_{\text{paarweise}} = 71.4\,\%$, so dass zunächst lediglich sechs Items zur Verfügung standen. Bei der Analyse der Expertenbeurteilungen der Items zu dieser Skala ist aufgefallen, dass sechs Items (i6.06, i6.09, i6.05, i6.11, i6.14, i6.16) zwar korrekt klassifiziert wurden, in Bezug auf den Typ des Arguments jedoch falsch zugeordnet wurden. Bei der inhaltlichen Analyse dieser Itemtexte ist aufgefallen, dass der Verweis auf bereits in das Experiment implementiertes Expertenwissen nur sehr implizit enthalten ist. Aus diesem Grund wurden Items umformuliert, so dass ein stärkere inhaltliche Passung zur Arbeitsdefinition vorliegt z. B. das Item i6.06 (vgl. Abschnitt B.2):

Bei meiner Entscheidung berücksichtige ich, dass das Experiment professionell aussah,

entsprechend umformuliert zu

Bei meiner Entscheidung berücksichtige ich, dass das Experiment professionell aussah, weil es von Fachleuten aufgebaut wurde.

Ein weiteres Item (i6.16) bezog sich stark auf das Computerexperiment („Entwickler des Experiments“) und wurde so umformuliert, dass es inhaltlich nicht von einer der beiden Subgruppen bevorzugt beantwortet werden kann.

Nach der inhaltlichen Beurteilung durch die Experten lagen daher 63 inhaltlich valide Items vor (18 Intuition, 12 Expertenwissen, 13 Messunsicherheiten, 20 Daten als Evidenz). Bei 81 % der Items erreichten die Experten eine Übereinstimmung von $P_{\text{paarweise}} = 75\%$

(identisches Urteil bei sieben von acht Experten) oder besser in beiden Testteilen. Bei lediglich sechs Items musste das abgeschwächte Kriterium von $P_{\text{paarweise}} = 71.4\,\%$ (identisches Urteil bei sechs von acht Experten) in maximal einem Testteil angewendet werden. Sechs Items der Skala Expertenwissen wurden inhaltlich überarbeitet und angepasst.

Zur abschließenden Bewertung wurde für die gesamte Itembatterie ein zufallskorrigiertes Übereinstimmungsmaß in Form von Fleiss' κ berechnet, wobei die sechs zuvor inhaltlich überarbeiteten Items der Skala Expertenwissen ausgeschlossen wurden. Auf Ebene des Gesamttests ergibt sich daher eine Beurteilerübereinstimmung von Fleiss' $\kappa = .83$ hinsichtlich der Beurteilung der Zentralität bzw. Peripherität der Items sowie ein Fleiss' $\kappa = .90$ bei der Einordnung der Items in die Argumentkategorien. Die Berechnung eines unjustierten Übereinstimmungsmaßes auf Skalenebene ist nicht sinnvoll, da sich die Randverteilungen nicht homogen verhalten. So ist z. B. die Grundwahrscheinlichkeit, dass ein Item der Skala Intuition in die Skala Daten als Evidenz eingeordnet wird, geringer als die Wahrscheinlichkeit, dass dieses Item in die Skala Intuition eingeordnet wird. Dieser „Effekt unterschiedlicher Grundwahrscheinlichkeiten" (Wirtz & Caspar, 2002, S. 64) führt dann zu einer systematischen Verzerrung der Koeffizienten (siehe auch Feinstein & Cicchetti, 1990).

B.5.4 *Diskussion*

Zur Beurteilung eines Expertenratings ist die Expertise der Beurteiler von hoher Bedeutung. Vor ihrem akademischen Hintergrund ist die Stichprobe der teilnehmenden Experten in der vorliegenden Arbeit als hinreichend geeignet für die Beurteilung der Itemtexte zu bewerten. Alle Experten konnten, nach einer textbasierten Schulung, die Testitems beurteilen. Bemerkenswert ist, dass auch Experten aus naturwissenschaftlichen Disziplinen abseits der Physik in der Lage sind, die Itemtexte problemlos zu beurteilen. Dies deutet auch auf eine hinreichend präzise und inhaltlich verständliche Beschreibung der Argumentkategorien hin.

Weiterhin ist anzumerken, dass zunächst ein sehr strenges Kriterium zur Beurteilung der inhaltlichen Validität eines Items gewählt wurde. Bei einer paarweisen prozentualen Übereinstimmung von $P = .75$, was einer Übereinstimmung von sieben von acht Ratern entspricht, wurde ein Item akzeptiert. Von diesem Kriterium wurde nur in sechs begründeten Einzelfällen abgewichen. Auf Itemebene kann daher davon ausgegangen werden, dass ein hohes Maß inhaltlicher Validität vorliegt.

Die abschließend bestimmte Übereinstimmung auf Skalenebene für die gesamte Itembatterie ist in beiden Textteilen ebenfalls als sehr gut

zu beurteilen, da Fleiss' κ einen Wert größer .75 annimmt (Wirtz & Caspar, 2002, S. 59).

Durch das Expertenrating konnten Items der Skala Expertenwissen identifiziert werden, die inhaltlich sehr allgemein formuliert waren. Auf diesen Mangel inhaltlicher Validität ist durch eine Überarbeitung der Itemtexte angemessen reagiert worden. Es ist jedoch anzumerken, dass eine erneute inhaltliche Beurteilung dieser sechs Items durch Experten nicht durchgeführt wurde.

Abschließend lässt sich konstatieren, dass eine Batterie aus 63 Items vorliegt, die eine hohe inhaltliche Validität aufweist. Dies begründet sich zum einen aus der Tatsache, dass die Itemtexte anhand von etablierten Instrumenten und aus Interviews entnommenen Schüleraussagen entwickelt wurden (siehe Abschnitt B.3). Zum anderen deuten darauf die Ergebnisse der Expertenbeurteilung hin, bei der die Items zur Selektion ein sehr strenges Kriterium erreichen mussten.

B.6 FESTLEGUNG EINES GEEIGNETEN ANTWORTFORMATS

Die vorliegenden Aussagen sollen als Likert-skalierte Items vorgelegt werden und durch Abfrage der Zustimmung zu bzw. Ablehnung der jeweiligen Aussage die Verwendung bestimmter Argumente im latenten Entscheidungsprozess messbar machen. In diesem Zusammenhang ist die Wahl eines geeigneten Antwortformates ein weiteres Entscheidungsfeld. Die Zahl der verwendeten Stufen und ihre Beschriftung ist in der Literatur intensiv diskutiert worden (für einen Überblick siehe Bühner, 2010). In der vorliegenden Arbeit wurde sich für eine fünfstufige Skala entschieden, die entsprechend des Vorschlags von Rohrmann (1978) mit „trifft gar nicht zu / trifft wenig zu / trifft teils/teils zu / trifft ziemlich zu / trifft völlig zu“ beschriftet ist. Für dieses Vorgehen sprechen zwei Gründe: Zum einen scheint die Bewertung der vorliegenden Itemtexte bei einer fünfstufigen Skala noch eindeutig möglich. Zum anderen verhalten sich die in nachfolgenden Analysen verwendeten ML-Verfahren im Rahmen der Schätzung von Strukturgleichungsmodellen und konfirmatorischen Faktorenanalysen bei Verwendung von mindestens fünfstufigen Likert-Skalen robust (für eine ausführliche Darstellung zu diesem Aspekt siehe Abschnitt 4.7.2).

B.7 EVALUATION DER PSYCHOMETRISCHEN QUALITÄT UND KONSTRUKTVALIDITÄT

Entsprechend des Entwicklungsziels 3 (vgl. Abschnitt B.1) wird im folgenden Abschnitt die Studie zur Evaluation der psychometrischen Qualität der Items und der Konstruktvalidität dargelegt.

B.7.1 *Methoden*

B.7.1.1 *Untersuchungsdesign*

Das Design der Untersuchung war identisch mit der Interviewstudie in Abschnitt 2.3, wobei an Stelle des Interviews nun die zu testende Itembatterie zu bearbeiten ist. Der Fragebogen erfasst zunächst einige Probandenmerkmale zur Charakterisierung der Stichprobe. Nach dem Aufstellen einer Hypothese werden die Probanden dann aufgefordert, die Hypothese im Experiment zu überprüfen und im Anschluss den Fragebogen zu bearbeiten. Der Fragebogen erfragt zunächst, ob die eingangs aufgestellte Hypothese aufgrund der experimentellen Daten verworfen bzw. beibehalten wird. Im weiteren Verlauf des Fragebogens werden die Probanden aufgefordert, die nachfolgenden Items im Hinblick auf die Bedeutung dieser Aussagen im persönlichen Entscheidungsprozess zum Beibehalten bzw. Verwerfen der eingangs aufgestellten Hypothese zu bewerten.

Der verwendete Fragebogen ist in Abschnitt B.8.2 dargestellt.

B.7.1.2 *Ablauf der Untersuchung*

Seitens der durchführenden Personen wurde Wert darauf gelegt, die teilnehmenden Schüler über Ziele und Inhalt der Untersuchung in Kenntnis zu setzen. Es wurde ferner über die Freiwilligkeit der Teilnahme und die Anonymität der Datenerfassung aufgeklärt, sowie deutlich dargelegt, dass es sich nicht um eine klassische Test- bzw. Leistungssituation handelt. Währenddessen wurden die Fragebögen ausgeteilt. Die Fragebögen ordneten durch eine Kennzeichnung die Probanden randomisiert einer der beiden Gruppen zu. Die erste Seite des Fragebogens wurde gemeinsam gelesen. Die Schüler bearbeiteten dann den Fragebogen in Stillarbeit bzw. führten das Experiment durch. Die Untersuchung dauerte etwa 60 Minuten. Abschließend wurde den teilnehmenden Schülern und den beteiligten Lehrkräften für die Teilnahme gedankt.

Während der Durchführung der Studie erstand der subjektive Eindruck, dass die teilnehmenden Schüler nur auf geringe Erfahrung im Umgang mit Experimenten zurückgreifen konnten.

B.7.1.3 *Datenaufbereitung und -analyse*

Die Daten wurden zunächst in eine tabellarische Form überführt. Alle nachfolgenden Analysen wurden innerhalb der freien Statistiksoftware R (R Core Team, 2014) durchgeführt. Zum Einsatz kamen ferner das Paket `psych` (Revelle, 2015) zur Berechnung gängiger psychometrischer Größen, die Pakete `lavaan` (Rosseel, 5 2012) und `semTools` (Pornprasertmanit et al., 2014) zur Berechnung der konfirmatorischen Faktorenanalysen, sowie `ggplot2` (Wickham & Chang, 2015) zur Erzeugung von Diagrammen. 19 Datensätze wurden als „Musterkreu-

zer“ identifiziert und für nachfolgende Analysen aus der Stichprobe entfernt. Von den verbleibenden 135 Datensätzen sind 116 Datensätze vollständig. Der maximale Anteil fehlender Werte pro Item ist für den gesamten Test kleiner als 5 %. Bezüglich des Umgangs mit den fehlenden Daten wird daher in den folgenden Analysen von der Annahme *missing at random* ausgegangen (Lüdtke et al., 2007; Rubin, 1976).

B.7.1.4 *Stichprobe*

An der Evaluation des Tests nahmen $n = 154$ Probanden einer Gemeinschaftsschule in Berlin teil. Die Stichprobe setzte sich aus vier Klassen der 8. Jahrgangsstufe, zwei der 9. Jahrgangsstufe und einer 10. Klasse zusammen. Die Tatsache, dass an der Studie eine 10. Klasse teilgenommen hat, ist einem Planungsfehler seitens der Schulleitung geschuldet. Die Stichprobe ist als Gelegenheitsstichprobe zu klassifizieren. Alle nachfolgenden Ergebnisse beziehen sich auf das verkleinerte Sample nach Ausschluss der Musterkreuzer. 45 % der Probanden waren weiblich, das Alter lag zwischen 13 und 17 Jahren und betrug im Mittel 14.54 Jahre. Da an der Schule je nach angestrebtem Schulabschluss nicht in allen Lerngruppen Noten vergeben werden, konnten lediglich 51 Probanden Auskunft über die Physiknote auf dem letzten Zeugnis geben, der Median liegt bei 3. 70 Probanden haben am Realexperiment gearbeitet, 65 Probanden am Computerexperiment.

B.7.2 *Ergebnisse*

Bei allen nachfolgenden Itemanalysen wird die vorliegende fünfstufige Ratingskala als intervallskaliert betrachtet, um parametrische Verfahren verwenden zu können. Dieses Vorgehen ist in der Literatur kontrovers diskutiert. Es zeigt sich jedoch, dass ein Großteil der statistischen Verfahren robust gegenüber Verletzungen des Skalenniveaus sind (z. B, Jamieson, 2004; Knapp, 1990; Marcus-Roberts & F. S. Roberts, 1987; Norman, 2010). In Abschnitt 4.7.2 wird diese Entscheidung bei der Verwendung der auf dem ML-Verfahren basierenden Methoden zusätzlich legitimiert.

B.7.2.1 *Deskriptivstatistische Itemanalyse*

Zunächst wurde die Itembatterie nach folgenden drei Kriterien der klassischen Testtheorie verkleinert:

1. Die Itemschwierigkeit P soll außerhalb extremer Werte liegen, d. h. $5 < P < 95$ (Pospeschill, 2010).

2. Die *part-whole* korrigierte Trennschärfe r der Items soll hinreichend groß sein, d. h. $r > .4$ (Moosbrugger & Kevala, 2008; Pospeschill, 2010).

Bei einer rein auf der Trennschärfe beruhenden Itemselektion besteht die Gefahr, dass das resultierende Instrument nicht in der Lage ist in den extremen Merkmalsausprägungen hinreichend zu differenzieren, da Unterschiede in den Schwierigkeitsindizes naturgemäß die Inter-Item-Korrelationen reduzieren. Der Selektionskennwert nach Lienert und Raatz (1998) stellt einen Versuch dar, Trennschärfe, Varianz und Itemschwierigkeit in ein integriertes Maß zu überführen. Der Selektionskennwert ergibt sich aus dem Quotienten der Trennschärfe r und der doppelten Standardabweichung des Items (Bühner, 2010; Pospeschill, 2010).

3. Aus den verbleibenden Items sollen diejenigen mit dem höchsten Selektionskennwert, jedoch maximal acht pro Argumentkategorie, ausgewählt werden.

Im Rahmen der Itemselektion wurde in besonderem Maße darauf geachtet, dass die Itemauswahl anhand psychometrischer Kenngrößen nicht zu einer inhaltlichen Einschränkung der zu messenden Konstrukte führt. Tabelle 25 stellt die relevanten Kenngrößen der nach den o. g. Kriterien ausgewählten Items dar. Entsprechend der Kriterien wurde die Itembatterie auf 31 Items reduziert. Die Schwierigkeiten der verbleibenden Items liegen durchweg in einem mittleren Bereich (d. h. $P_i \approx 50$), überstreichen dabei jedoch immer ein gewisses Intervall (Intuition: $25.8 \leq P_i \leq 46.5$; Expertenwissen $48.3 \leq P_i \leq 58.5$; Messunsicherheiten $43.3 \leq P_i \leq 51.3$; Daten als Evidenz $60.5 \leq P_i \leq 73.8$). Die Trennschärfe der Items liegt entsprechend des Kriteriums bei $r > .4$. Die Items von drei der vier Skalen erreichen jedoch deutlich höhere Werte (Intuition: $.52 \leq r_i \leq .76$; Expertenwissen $.68 \leq r_i \leq 78$; Messunsicherheiten $.40 \leq r_i \leq .53$; Daten als Evidenz $.53 \leq r_i \leq .70$).

B.7.2.2 *Testung der faktoriellen Struktur mittels konfirmatorischer Faktorenanalysen*

Zur Evaluation der Konstruktvalidität eines psychometrischen Tests werden im allgemeinen theoretische Annahmen mittels empirischer Daten überprüft.

Bei dem vorliegenden Testentwurf ist von einer zugrundeliegenden vierfaktoriellen Struktur auszugehen, da vier Argumentkategorien operationalisiert wurden. Diese vierfaktorielle Struktur soll durch eine konfirmatorische Faktorenanalyse (CFA) überprüft werden. Neben Verfahren der Item-Response-Theorie (IRT) stellt die CFA die einzige Möglichkeit dar, insbesondere die faktorielle Validität, d. h. eine dem Test zugrundeliegende Dimensionalität, inferenzstatistisch zu überprüfen (Brown, 2006; Hartig et al., 2012). Die CFA schätzt anhand eines a priori spezifizierten Modells (hier: das vierfaktorielle System der Argumentkategorien) eine modellkonforme Populationskovarianzmatrix und prüft, ob diese bedeutsam von der Stichprobenkovari-

anzmatrix abweicht. Dadurch kann Aufschluss über die Passung des Modells (in diesem Fall die spezifizierte Zuordnung zwischen Item und entsprechender Argumentkategorie) und somit ein Hinweis auf das Vorliegen von Konstruktvalidität bzw. faktorieller Validität gewonnen werden (Brown, 2006; Eid et al., 2013). Ferner kann die diskriminante Validität des Tests untersucht werden, indem innerhalb des in der CFA spezifizierten Strukturmodells Korrelationen zwischen den latenten Faktoren geschätzt werden. Das Vorliegen von konvergenter Validität kann wiederum durch die Interpretation der Faktorladungen abgeschätzt werden (Brown, 2006).

Im Gegensatz zur explorativen Faktorenanalyse, bei der vom Forscher lediglich die Zahl der Faktoren bestimmt werden kann, ist die Spezifikation eines CFA-Modells stark theoriegeleitet: Der Forscher kann neben der Zahl der Faktoren auch die Struktur zwischen Indikatoren und latenten Faktoren festlegen sowie die Nebenladungen zu null fixieren. Es wird daher ein deutlich sparsameres Modell auf Passung zur empirischen Datenlage getestet. Die CFA bietet ferner die Möglichkeit, die Residuen der Indikatoren, d. h. den Teil der Indikatorvarianz, der nicht durch den latenten Faktor erklärt werden kann, zu modellieren (so können z. B. korrelierte Fehlerterme eingefügt werden die aufgrund von Methodeneffekten entstehen, z. B. bei der Verwendung ähnlicher Wörter in den Itemtexten). Brown (2006) empfiehlt daher die Verwendung von CFAs, wenn a priori Hypothesen über Struktur und Zusammenhänge des Messmodells bestehen (siehe auch Henson & J. K. Roberts, 2006). Die Verwendung von konfirmatorischen Faktorenanalysen bietet außerdem die Möglichkeit, die Äquivalenz von Messmodellen (Messinvarianz) über verschiedene Subgruppen zu evaluieren, sowie in diesem Kontext latente Mittelwertsunterschiede zu prüfen (vgl. Abschnitt 5.4). CFAs können außerdem leicht zu Strukturgleichungsmodellen erweitert werden, indem strukturelle Beziehungen zwischen latenten Variablen spezifiziert und untersucht werden, was insbesondere zur Beantwortung der Hauptforschungsfragen von Bedeutung ist Abschnitt 3.2. Eine detailliertere Diskussion der Vorteil von konfirmatorischen Faktorenanalysen liefert z. B. Brown (2006, S. 51).

Die Modellpassung kann anhand verschiedener Gütemaße evaluiert werden. Eine Übersicht sowie eine Festlegung auf gängige Cutoff-Kriterien findet sich in Abschnitt E.1.

AUSWAHL DES SCHÄTZVERFAHRENS Die Parameter wurden im Rahmen der hier gerechneten konfirmatorischen Faktorenanalysen wie auch in der Hauptuntersuchung der vorliegenden Arbeit durch robuste ML-Verfahren geschätzt. Für die vorliegenden Daten zeigt der Test nach Mardia (1970), dass die Nullhypothese der multivariaten Normalverteilung verworfen werden muss (Schiefe $S = 233.3$, $p_S < .001$, Kurtosis $K = 3.96$, $p_K < .001$; Kriterien: $S < 2$; $K < 7$, siehe

West, Curran und Finch, 1996). Robuste ML-Verfahren liefern bei ordinalen Indikatoren mit fünf oder mehr Stufen auch bei Verletzung der Verteilungsvoraussetzungen unverzerrte Parameterschätzungen und Standardfehler (Pui-Wa & Qiong, 2012; Rhemtulla et al., 2012). Eine tiefergehende Argumentation für die Auswahl des Schätzverfahrens findet sich im Methodenteil des Haupttextes (Abschnitt 4.7.2).

SPEZIFIZIERUNG DER ZU TESTENDEN MODELLE Im Folgenden werden zunächst vier zu testende Modelle spezifiziert, die im Rahmen der CFAs geprüft werden sollen. Bei den Modellen 1 bis 3 handelt es sich um Modelle mit der theoretisch zu erwartenden vierfaktoriellen Struktur. Anhand des Modells 1 wurde zunächst die Itembatterie weiter verkleinert. Modell 2 und 3 unterscheiden sich lediglich in der Struktur der latenten Variablen und Modell 4 stellt eine theoretisch denkbare Alternative dar. Im Folgenden sollen diese Modelle detaillierter beschrieben werden.

MODELL 1: 4-FAKTORIELLES MODELL, 31 ITEMS Die in diesem Modell zugrundeliegenden vier latenten Faktoren entsprechen den operationalisierten vier Argumentkategorien. Es soll geprüft werden, ob die Zuordnung der Items, wie durch die Itementwicklung beabsichtigt bzw. durch die Expertenstudie im Hinblick auf die inhaltliche Validität geprüft wurde, auch empirisch zu beobachten ist. Dieses Modell soll außerdem herangezogen werden, um die nach der deskriptivstatistischen Analyse vorliegende Itembatterie von 31 Items weiter zu verringern. Im Rahmen der konfirmatorischen Faktorenanalyse können die standardisierten Diskriminationsparameter (d. h. die Faktorladungen) zur Itemselektion herangezogen werden (Brown, 2006; Eid et al., 2013). Aus pragmatischen Gründen soll die Itembatterie auf fünf Items pro Kategorie verringert werden. Unter Berücksichtigung der inhaltlichen Breite der Konstrukte sollen je Kategorie die fünf Items mit dem höchsten Diskriminationsparameter aus dem Messmodell extrahiert werden.

MODELL 2: 4-FAKTORIELLES MODELL, 5 ITEMS PRO FAKTOR, LATENTE FAKTOREN OBLIQUE Im Gegensatz zu Modell 1 beinhaltet dieses Modell lediglich fünf Items pro Kategorie, ausgewählt nach der Höhe der Diskriminationsparameter in Modell 1. Im Strukturteil des Modells ist ferner Kovariation zwischen den latenten Faktoren zugelassen, um die Struktur der latenten Faktoren zu testen.

MODELL 3: 4-FAKTORIELLES MODELL, 5 ITEMS PRO FAKTOR, LATENTE FAKTOREN ORTHOGONAL Identisch zu Modell 2, jedoch ist die Kovariation der latenten Faktoren hier zu null spezifiziert.

MODELL 4: 1-FAKTORIELLES MODELL, 5 ITEMS PRO FAKTOR Da es durchaus denkbar ist, dass den Antwortreaktionen der Probanden auf die vorgelegten Items lediglich ein Faktor zugrundeliegt (der z. B. im Sinne einer allgemeinen Antworttendenz zu interpretieren sein könnte), wird ein Modell spezifiziert, bei dem nur ein Faktor zur Erklärung der Itemvarianz herangezogen wird.

SKALIERUNG DER LATENTEN VARIABLEN Zur Festlegung der Metrik der latenten Variablen wird im Allgemeinen vorgeschlagen, die Faktorladung des Indikators mit dem stärksten Zusammenhang zum Konstrukt auf 1 zu fixieren. Dieser *marker indicator*-Ansatz (Brown, 2006, S. 106) ist jedoch nur dann gerechtfertigt, wenn diese Entscheidung argumentativ begründet werden kann. Da in diesem Stadium der Testentwicklung jedoch keine begründete Aussage über die Item-Faktor-Beziehungen getroffen werden kann, werden die Varianzen der latenten Faktoren auf 1 fixiert, um die Metrik der latenten Variablen zu bestimmen.

TESTUNG DER MODELLE Die Gütemaße der Modelltestungen 1 bis 4 sind in Tabelle 24 dargestellt. Bei der Beurteilung der Modelle ist zunächst festzustellen, dass bei allen Modellen der χ^2-Test des exakten Modelltests signifikant ausfällt. Wie jedoch in Abschnitt E.1 diskutiert, wird dieses Maß nicht interpretiert. Modell 1 weist einen schwach akzeptablen Modellfit auf. Aus diesem Modell wurden anhand der vollständig standardisierten Faktorladungen (Diskriminationsparameter) pro Skala die fünf Items mit dem höchsten Diskriminationsparametern ausgewählt, um die Itembatterie weiter zu verkleinern. Die Diskriminationsparameter der Modelle sind in Tabelle 25 dargestellt. Das resultierende Modell 2 weist einen guten Modellfit auf, da alle Kriterien unterhalb der in Abschnitt E.1 festgelegten Cut-off-Wertes liegen und der CFI mit .94 näherungsweise einen guten Modellfit vermuten lässt. Das unmittelbar konkurrierende Modell 3, das sich zu Modell 2 nur dahingehend unterscheidet, dass Korrelation zwischen den Faktoren nicht erlaubt ist, sowie das Modell 4 müssen hingegen abgelehnt werden: Insbesondere SRMR und RMSEA liegen über den Cut-off-Kriterien.

Tabelle 24: Gütemaße konkurrierender Messmodelle des Testentwurfs. In Modell 1 laden alle Items auf den entsprechenden Faktoren. Anhand dieses Modells wurden pro Faktor die 5 geeignesten Items ausgewählt und erneut getestet (Modell 2). Im Gegensatz dazu lässt Modell 3 keine Kovarianz zwischen den latenten Faktoren zu. Modell 4 untersucht die Passung bei Zuordnung aller Items auf lediglich einen allgemeinen Faktor. Nach Beurteilung der Gütemaße weist das Modell 2 einen den Alternativmodellen überlegene Passung auf.

#	Modell	χ^2	df	p	χ^2/df	CFI	RMSEA [90 %C.I.]	$p_{\text{RMSEA}<.05}$	SRMR
1	4-faktorielles Modell, alle Items	599.5	428	.00	1.40	.88	.05 [.04; .06]	.23	.08
2	4-faktorielles Modell, 5 Items	216.4	164	.00	1.32	.94	.05 [.03; .06]	.54	.07
3	4-faktorielles Modell, 5 Items, orthogonal	258.6	170	.00	1.52	.90	.06 [.05; .08]	.09	.13
4	1-faktorielles Modell, 5 Items	818.4	170	.00	4.81	.29	.17 [.16; .18]	.00	.19

Hinweise: CFI ... Comparative-Fit-Index; RMSEA ... Root Mean Square Error of Approximation; SRMR ... Standardized Root Mean Square Residual

Tabelle 25: Psychometrische Kennwerte der Items des Testentwurfs. Die Tabelle enthält nur die nach den Ergebnissen der deskriptivstatistischen Itemanalyse ausgewählten Items. Kodierung der Likert-Stufen von 0 bis 4. Die Itemschwierigkeit bzw. der Popularitätsindex P (Lienert & Raatz, 1998) wird umso größer, desto häufiger ein Item im Sinne des Merkmals beantwortet wird. Die beiden letzten Spalten berichten die Ergebnisse der CFA-basierten Itemanalyse. Alle berichteten Diskriminationsparameter sind signifikant von null verschieden.

		Deskriptiv-statistische Itemanalyse					CFA -modellbasierte Itemanalyse	
Kategorie	Item	r	M	SD	P	SK	$\lambda_{\text{Modell 1}}$	$\lambda_{\text{Modell 2}}$
Intuition	i1.03	.69	1.4	1.2	34.9	.28	.77*	.76
	i1.09	.68	1.6	1.3	38.8	.27	.70*	.68
	i1.11	.57	1.8	1.2	43.7	.25	.63	
	i1.15	.52	1.0	1.2	25.8	.22	.58	
	i1.16	.72	1.5	1.4	36.2	.26	.79*	.80
	i1.18	.62	1.9	1.4	46.5	.22	.66*	.65
	i1.20	.53	1.5	1.3	37.9	.20	.57	
	i1.21	.76	1.5	1.3	36.6	.30	.83*	.83
Expertenwissen	i6.05	.78	2.1	1.2	52.9	.31	.81*	.81
	i6.06	.70	1.9	1.3	48.3	.27	.72	
	i6.08	.78	2.3	1.3	57.5	.29	.82*	.84
	i6.11	.73	2.3	1.3	57.4	.29	.78*	.80
	i6.14	.68	2.2	1.4	53.9	.25	.71	
	i6.15	.77	2.3	1.4	58.5	.29	.80*	.77
	i6.16	.75	2.2	1.3	55.5	.28	.80*	.81
Messunsicherheiten	i7b.04	.52	2.1	1.1	43.3	.23	.64*	.76
	i7b.05	.47	1.8	1.3	44.1	.18	.50	
	i7b.07	.41	1.7	1.1	46.3	.18	.46	
	i7b.08	.47	1.9	1.2	56.5	.20	.54*	.46
	i7b.10	.40	1.9	1.1	48.3	.18	.44	
	i7b.16	.50	2.0	1.1	50.4	.22	.57*	.46
	i7b.17	.53	1.8	1.2	43.9	.21	.58*	.48
	i7b.22	.49	2.3	1.2	51.3	.20	.61*	.73
Daten als Evidenz	i8.05	.70	2.8	1.1	70.1	.31	.78*	.76
	i8.09	.64	2.4	1.1	60.4	.29	.69*	.69
	i8.11	.55	2.6	1.1	65.9	.25	.64*	.68
	i8.14	.56	2.8	1.2	70.8	.24	.62*	.66
	i8.16	.54	3.0	1.2	73.7	.23	.58	
	i8.18	.64	2.6	1.2	64.0	.26	.71*	.70
	i8.22	.53	2.9	1.1	73.3	.23	.55	
	i8.25	.55	2.6	1.1	66.1	.25	.59	

Hinweise: r ... korrigierte Item-Skala-Korrelation; M ... Mittelwert; SD ... Standardabweichung; P ... Itemschwierigkeit; SK ... Selektionskennwert; $\lambda_{\text{Modell 1}}$ ... standardisierter Diskriminationsparameter aus Modell 1; $\lambda_{\text{Modell 1}}$... standardisierter Diskriminationsparameter aus Modell 2; *... Item ausgewählt für Modell 2

Tabelle 26: Modell-implizierte Reliabilitäten ρ der Subskalen

Skala	ρ
Intuition	.86
Expertenwissen	.90
Messunsicherheiten (explizit)	.72
Daten als Evidenz	.82

Da das Modell 2 die beste Passung zu den Daten aufweist, werden alle weiteren Analysen, insbesondere die Analysen zur Messinvarianz anhand dieses Modells vorgenommen. Die Diskriminationsparameter der Items weisen überwiegend hohe bis sehr hohe Werte auf ($\bar{\lambda} = .71, SD = .10, \lambda_{\text{min}} = .50, \lambda_{\text{max}} = .84$). Das Quadrat der standardisierten Diskriminationsparameter kann als der durch den latenten Faktor aufgeklärten Varianzanteil R^2 interpretiert werden (Brown, 2006). Für z. B. Item `i1.21` folgt daher $\lambda^2 = .83^2 = .69$, d. h. 69 % der Indikatorvarianz werden durch den latenten Faktor erklärt.

Modell 2 lässt Kovariation zwischen den latenten Faktoren zu: Statistisch signifikant von null verschieden sind lediglich die latenten Korrelationen zwischen den Faktoren Intuition und Daten als Evidenz ($r_{\text{lat}} = -.50$, $p < .001$), zwischen Expertenwissen und Messunsicherheiten ($r_{\text{lat}} = -.34$, $p = .01$), sowie zwischen Expertenwissen und Daten als Evidenz ($r_{\text{lat}} = .23$, $p = .04$).

Die modell-implizierten Reliabilitäten der vier Skalen sind in Tabelle 26 dargestellt. Die Reliabilitäten der Subskalen liegen im Intervall von $.72 < \rho < .90$.

TESTUNG DER MESSINVARIANZ Die Notwendigkeit der Überprüfung der Messinvarianz des Testentwurfs begründet sich aus der Tatsache, dass der Test nur dann sinnvoll bearbeitet werden kann, wenn ihm eine Experimentiersituation vorausgeht (für eine ausführliche Begründung siehe Abschnitt B.1). Da in der vorliegenden Arbeit u. a. der Einfluss verschiedener Merkmale von Experimentiersituationen auf bestimmte latente Variablen geprüft werden soll, ist daher zunächst sicherzustellen, dass in Stichproben, die mit verschiedenen Experimenten gearbeitet haben, überhaupt identische Konstrukte vorliegen und vergleichbar gemessen werden können.

In diesem Kapitel ist das Vorgehen zur Evaluation der Messinvarianzbedingungen des Testentwurfs dokumentiert. Da es sich dabei nicht um ein Standardverfahren in der naturwissenschaftsdidaktischen Forschung handelt, wird zunächst eine methodische Einführung in das Verfahren gegeben. Es wird sich dabei an dem Standard-

werk von Brown (2006) orientiert. Kline (2016, S. 398) bietet ebenfalls eine gute Einführung, verwendet jedoch eine andere Nomenklatur. Verschiedene Ausprägungen der Messinvarianz (Messäquivalenz) werden nicht konsistent benannt. Während Brown (2006) von konfiguraler (Gleichheit der Faktorladungsmatrix), metrischer (Gleichheit der Faktorladungen) und skalarer (Gleichheit der Item-Intercepts) spricht, sind diese Ausprägungen bei Kline (2016) mit konfiguraler, schwacher und starker Invarianz benannt.

Im Rahmen der konfirmatorischen Faktorenanalyse kann die Äquivalenz von Testparametern über Teilgruppen einer Stichprobe hinweg analysiert werden. Es werden dabei sukzessive Parameter des Messmodells (Faktorladungen und -struktur, Intercepts, Residualvarianzen) und des Strukturmodells (d. h. Faktorvarianzen, -kovarianzen und latente Mittelwerte) auf Äquivalenz in den zu vergleichenden Gruppen analysiert. Die Untersuchung dieser Parameter im Messmodell evaluiert das Vorliegen von Messinvarianz, die Untersuchung der Übereinstimmung von Parametern im Strukturmodell, z. B. auf Gleichheit der latenten Mittelwerte, gibt Aufschluss über Populationsunterschiede, wobei im Gegensatz zu analogen Verfahren wie *t*-Test oder ANOVA der Messfehler der Indikatoren explizit modelliert wird (Brown, 2006). In der Multigruppen-CFA wird eine Faktorenanalyse in zwei oder mehreren Gruppen simultan durchgeführt, wobei die Messinvarianz durch sukzessive Beschränkungen im Messmodell evaluiert werden kann.

Entsprechend der zu prüfenden Parameterrestriktionen existieren verschiedene Formen der Invarianz, die in Multigruppen-CFAs sukzessive getestet werden können: In einem ersten Schritt wird zunächst die Struktur der Faktorladungsmatrizen, d. h. die Zuordnung der Indikatoren zu den latenten Faktoren, in den Gruppen geprüft. Die Höhe der Faktorladungen kann dabei in allen Gruppen zunächst frei variieren. Technisch werden dabei die gruppenspezifischen Stichprobenkovarianzmatrizen mit der modellkonformen Kovarianzmatrix verglichen. Der Fit des Multigruppen-Modells wird dabei jedoch global, d. h. gemeinsam für alle Gruppen evaluiert. Deuten die Gütemaße auf einen nicht akzeptablen Fit hin, ist die Faktor-Item-Struktur zwischen den Gruppen verschieden und folglich die Vergleichbarkeit der Konstrukte über die Gruppen hinweg nicht gegeben. Weitere Invarianztests sind dann nicht sinnvoll und Gruppenvergleiche folglich nicht legitim. Kann das Modell akzeptiert werden, liegt *konfigurale Invarianz* vor. Es kann dann davon ausgegangen werden, dass die Indikatoren in beiden Gruppen dieselben latenten Faktoren messen. Es kann dann die nächst strengere Invarianzbedingung getestet werden. Beim Test auf *metrische Invarianz* wird die Annahme der Gleichheit der Faktorladungen über die Gruppen hinweg evaluiert. Dazu werden die Faktorladungen auf einen in beiden Gruppen identischen, aber frei zu schätzenden Wert fixiert. Weist dieses Modell gegenüber dem Mo-

dell der konfiguralen Invarianz keinen signifikant schlechteren Modellfit auf (zur Festlegung geeigneter Kriterien beim Vergleich hierarchisch geschachtelter Modelle siehe Anhang Abschnitt E.1.2), kann die Annahme der metrischen Invarianz aufrecht erhalten werden. Die Stärke der Beziehung zwischen manifesten Indikatoren und latenten Variablen ist daher in beiden Gruppen als gleich zu betrachten. Obwohl somit die Einheiten der Messung („unit of measurement", Byrne und Stewart, 2006, S. 296) und damit eine in allen Gruppen identische Veränderung eines latenten Faktors zu einer in beiden Gruppen gleichen Veränderung der Indikatoren führt (Temme & Hildebrandt, 2008), ist der Ursprung der Skalen (d. h. die Item-Intercepts) nicht zwingend identisch (Byrne & Stewart, 2006). Diese Annahme der *skalaren Invarianz* wird überprüft, indem das Modell um die Gleichheit der Item-Intercepts erweitert wird. Verschlechtern diese Restriktionen das Modell nicht signifikant, kann die Annahme skalarer Invarianz nicht verworfen werden; es können sowohl latente Mittelwerte als auch strukturelle Beziehungen der latenten Faktoren verglichen werden (Brown, 2006; Temme & Hildebrandt, 2008). In einem weiteren Schritt kann die Gleichheit der Residualvarianzen in beiden Gruppen getestet werden. Dieses sog. Niveau *strikter Invarianz* ist jedoch für den Vergleich latenter Mittelwerte nicht nötig und wurde in der vorliegenden Arbeit nicht getestet (Brown, 2006).

Übergeordnetes Ziel der folgenden Analysen ist die Evaluation der psychometrischen Qualität der Items, insbesondere im Hinblick auf die strukturellen Beziehungen zwischen manifesten Indikatoren und latenten Faktoren. Durch die Anwendung der Multigruppen-CFA erhöht sich jedoch die Zahl der freien Parameter fast um einen der Zahl der Gruppen entsprechenden Faktor. Vor diesem Hintergrund erscheint es sinnvoll, die Modelle zu verkleinern, um den Voraussetzungen des ML-Schätzverfahrens gerecht zu werden. Es ist daher notwendig, das bereits getestete vierfaktorielle Gesamtmodell in vier Teilmodelle zu zerlegen, sodass die Skalen zur Erfassung der einzelnen Argumentkategorien getrennt getestet werden können. Eine Legitimation für dieses Vorgehen lässt sich aus den folgenden Argumenten ableiten (A. Hildebrandt, persönliche Kommunikation, August 2013): a) Das zu favorisierende Modell 2 beschreibt lediglich zwischen je zwei Paaren latenter Faktoren eine signifikant von null verschiedene Korrelation. Starke strukturelle Zusammenhänge zwischen den Faktoren liegen daher nicht vor. b) Die theoretisch postulierten Beziehungen zwischen Indikatoren und den vier latenten Faktoren wurden in der Gesamtgruppe bereits bestätigt. Es ist nun nicht notwendig, dies in den einzeln Gruppen erneut zu tun. c) Die übergeordneten Hauptforschungsfragen sehen keine gruppenspezifische Evaluation der Faktorkovarianz vor.

Tabelle 27: Gruppenspezifischer Modellfit für die vier Submodelle. Im ersten Schritt der Messinvarianzanalyse wird geprüft ob die Messmodelle auch einzeln in den Teilstichproben eine akzeptable Passung aufweisen. Für jede der denkbaren Gruppen (R, V, Gesamt) wurde daher je Faktor ein Messmodell bestimmt. Die Tabelle zeigt die Modellgütemaße.

								90 % C.I.			
Fak	Gruppe	χ^2	df	p	χ^2/df	CFI	RMSEA	lower	upper	$p_{\text{RMSEA}<.05}$	SRMR
int	R	9.5	5	.09	1.91	.96	.12	.00	.22	.14	.05
	V	1.3	5	.94	0.25	1.00	.00	.00	.00	.98	.02
	Gesamt	5.1	5	.41	1.02	1.00	.01	.00	.11	.62	.03
exp	R	0.4	5	1.00	0.08	1.00	.00	.00	.00	1.00	.01
	V	6.4	5	.27	1.28	.99	.07	.00	.18	.36	.03
	Gesamt	3.5	5	.63	0.69	1.00	.00	.00	.07	.87	.02
mu	R	2.4	4	.66	0.60	1.00	.00	.00	.14	.73	.03
	V	3.8	4	.44	0.94	1.00	.00	.00	.18	.52	.04
	Gesamt	1.4	4	.85	0.34	1.00	.00	.00	.08	.90	.01
evi	R	3.4	5	.63	0.69	1.00	.00	.00	.11	.79	.03
	V	8.8	5	.12	1.77	.95	.11	.00	.21	.17	.05
	Gesamt	7.3	5	.20	1.45	.98	.06	.00	.13	.37	.03

Hinweise: CFI ... Comparative-Fit-Index; RMSEA ... Root Mean Square Error of Approximation; SRMR ... Standardized Root Mean Square Residual

Die Evaluation der Messinvarianz in der vorliegenden Arbeit orientiert sich an einem u. a. von Brown (2006, S. 270) vorgeschlagenen Ablauf:

1. Testung der CFA-Modelle separat in jeder einzelnen Gruppe
2. Test auf konfigurale Invarianz
3. Test auf metrische Invarianz
4. Test auf skalare Invarianz

Das Erreichen vorhergehender Niveaustufen der Messinvarianz versteht sich dabei als notwendige Voraussetzung für alle weiteren Analysen.

GRUPPENSPEZIFISCHE MODELLGÜTE In einem ersten Schritt werden die vier einfaktoriellen Sub-Modelle getrennt für jede Teilstichprobe evaluiert, d. h. es wird je ein Modell für jede Experimentiersituation gerechnet sowie ein Modell, dass die gesamte Stichprobe umfasst. Die Gütemaße dieser gruppenspezifischen Modellfits sind in Tabelle 27 dargestellt.

Da die Analyse für die weitere Prüfung der Messinvarianz relevant ist, werden im Folgenden die Modelle bewertet: Der χ^2-Test zeigt für kein Modell eine signifikante Abweichung von der Populationskovarianzmatrix an, der CFI weist für alle Fits gute bis hervorragende Werte auf. Der RMSEA zeigt für das Modell `int-R` und `evi-V` Werte $> .06$, wobei jedoch die untere Grenze des Konfidenzintervalls die null mit einschließt, außerdem ist der Signifikanztest für H_0 : RMSEA $< .05$ (Spalte $p_{\text{RMSEA}< .05}$) für alle Sub-Modelle $> .05$. Das Überschreiten des RMSEA sollte in diesem Fall nicht überinterpretiert werden: Es existieren eine Reihe von Hinweisen darauf, dass der RMSEA bei einer kleinen Anzahl an Freiheitsgraden (hier df = 5) nicht performant ist (z. B. Kenny, 2014). Alle der hier getesteten Sub-Modelle werden zur weiteren Evaluation der Messinvarianzbedingungen herangezogen.

MODELLFITS DER HIERARCHISCH GESCHACHTELTEN MODELLE ZUR PRÜFUNG DER MESSINVARIANZ Entsprechend des von Brown (2006) vorgeschlagenen Vorgehens wird im Folgenden sukzessive das Vorliegen von konfiguraler, metrischer und skalarer Invarianz geprüft. Die Ergebnisse dieser Modellschätzungen sind getrennt für jeden latenten Faktor in Tabelle 28 auf Seite 174 dargestellt. Für alle Modelle zeigt der χ^2-Test keine signifikante Abweichung an, der CFI und RMSEA nehmen gute bis sehr gute Werte an. Für alle konkurrierenden Modellvergleiche sind zum einen χ^2-Differenzentests gerechnet worden. Da dieser bei der Evaluation der Messinvarianz sensitiv reagiert, wurde zusätzlich der ΔCFI bestimmt (die Kriterien für den Vergleich hierarchisch geschachtelter Modelle sind in Abschnitt E.1.2).

Der χ^2-Differenzentest zeigt bei allen Skalen und über alle Stufen der Invarianz hinweg keine Verschlechterung der Modellpassung an. Außer beim Vergleich der Modelle konfiguraler bzw. metrischer Messinvarianz für den Faktor Daten als Evidenz sind alle ΔCFI-Werte positiv oder null. Die Verschlechterung des ΔCFIs liegt hier jedoch unter dem Cut-off-Kriterium $\Delta\text{CFI} < -.005$.

Tabelle 28: Modellfits der hierarchisch geschachtelten Modelle zur Prüfung der Messinvarianz. Für jede Argumentkategorie wurden die Messinvarianzbedingungen (konfigurale, metrische, skalare Messinvarianz) entsprechend des bei Brown (2006) beschriebenen Reihenfolge separat evaluiert. Modellvergleiche sind in den letzten drei Spalten dargestellt (Zur Beurteilung hierarchisch geschachtelter Modelle siehe Abschnitt E.1.2).

Fak	Niveau	χ^2	df	p	CFI	RMSEA	90 % C.I.	p	SRMR	$\Delta\chi^2$	p	ΔCFI
int												
	konfigural	9.55	10	.48	1.000	.00	.00 – .12	.64	.03			
	metrisch	12.88	14	.54	1.000	.00	.00 – .10	.71	.06			
	skalar	17.78	18	.47	1.000	.00	.00 – .10	.66	.06			
					konfigural vs. metrisch					3.21	.47	.000
					metrisch vs. skalar					5.05	.28	.000
exp												
	konfigural	5.18	10	.88	1.000	.00	.00 – .02	.98	.02			
	metrisch	8.74	14	.85	1.000	.00	.00 – .05	.96	.05			
	skalar	12.38	18	.83	1.000	.00	.00 – .05	.95	.06			
					konfigural vs. metrisch					4.60	.33	.000
					metrisch vs. skalar					4.03	.40	.000
mu												
	konfigural	6.10	8	.64	1.000	.00	.00 – .12	.73	.03			
	metrisch	11.80	12	.46	1.000	.00	.00 – .12	.61	.06			
	skalar	13.20	16	.66	1.000	.00	.00 – .09	.79	.06			
					konfigural vs. metrisch					5.44	.25	.000
					metrisch vs. skalar					1.07	.90	.000
evi												
	konfigural	11.45	10	.32	.989	.05	.00 – .13	.47	.04			
	metrisch	15.87	14	.32	.985	.04	.00 – .12	.49	.07			
	skalar	17.17	18	.51	1.000	.00	.00 – .10	.71	.07			
					konfigural vs. metrisch					4.40	.35	−.003
					metrisch vs. skalar					0.64	.96	.015

Hinweise: CFI ... Comparative-Fit-Index; RMSEA ... Root Mean Square Error of Approximation; SRMR ... Standardized Root Mean Square Residual

B.7.3 *Diskussion*

Auf den vorangegangenen Seiten wurde das Vorgehen und die Ergebnisse der Evaluation der psychometrischen Qualität der neu entwickelten Items berichtet. Entsprechend der in den Entwicklungszielen der Studie definierten Kriterien (Ziel 3, Abschnitt B.1) sollen die Ergebnisse der Analysen im Folgenden diskutiert werden:

Aus dem Set der inhaltlich validierten Items konnten anhand der klassischen Kriterien ingesamt 31 Items für weitere Analysen extrahiert werden (8 für die Kategorie Intuition, 7 für die Kategorie Expertenwissen, 8 für die Kategorie Messunsicherheiten (explizit), 8 für die Kategorie Daten als Evidenz). Dabei wurden zunächst Items mit extremen Schwierigkeiten ausgeschlossen. Die Schwierigkeiten der Items liegen durchweg in einem mittleren Bereich. Nach Pospeschill (2010) bedeuten mittlere Itemschwierigkeiten eine maximale Streuung der Itemantworten und eine hohe Differenzierung zwischen Probanden mit mittlerer Merkmalsausprägung. Es muss jedoch bemerkt werden, dass der Test nicht besonders gut bei Probanden in den Extremen der Merkmalsverteilung differenzieren kann. Es ist festzustellen, dass sich die Schwierigkeitsintervalle einzelner Skalen nicht überdecken (z.B Daten als Evidenz und Expertenwissen). Obwohl die einzelnen Skalen aufgrund des identischen Antwortformats vermutlich gleiche Messeinheiten besitzen, ist nicht sichergestellt, dass auch der Ursprung der Skalen identisch ist. Eine Interpretation dieses Ergebnisses ist daher nicht gestattet.

Um die Batterie weiter zu verkleinern, wurden Items mit einer klassischen Trennschärfe von $r < .4$ ausgeschlossen. Die Items von drei der vier Skalen erreichen jedoch deutlich höhere Werte. Es ist daher davon auszugehen, dass die Items das jeweilige Merkmal homogen gegenüber dem Gesamttest messen. Probanden mit niedriger Merkmalsausprägung werden entsprechend gering auf einem einzelnen Item abschneiden, während Probanden, deren Argumentation die entsprechende Kategorie berücksichtigt, hoch auf dem Item abschneiden. Die Trennschärfe der Testteile ist abschließend daher mit „angemessen“ bis „ausgezeichnet“ zu bewerten (Pospeschill, 2010).

Die Itembatterie wurde anhand des Selektionskennwerts weiter verkleinert. Dieses Vorgehen stellt sicher, dass eine rein auf der Trennschärfe basierende Itemauswahl das Intervall der Itemschwierigkeiten nicht einschränkt, da Varianz in den Schwierigkeitsparametern die Item-Skala-Korrelationen reduziert (Pospeschill, 2010).

Die Itemselektion folgte bis zu diesem Punkt anerkannten deskriptivstatistischen Kriterien der Itemselektion (Bühner, 2010; Lienert & Raatz, 1998; Pospeschill, 2010). Nach diesem Vorgehen konnten 31 Items extrahiert werden, von denen daher anzunehmen ist, dass sie klassischen psychometrischen Qualitätsmerkmalen entsprechen.

Zur Prüfung der zugrundeliegenden latenten Struktur des Testentwurfs wurde das Verfahren der konfirmatorischen Faktorenanalyse angewendet (Brown, 2006). Da eine weit entwickelte Theorie zur Struktur der Items vorlag, ist diese strukturprüfende Methode z. B. der explorativen Faktorenanalyse vorzuziehen (Brown, 2006; Henson & J. K. Roberts, 2006). Es wurde ein Messmodell aufgestellt, dass der postulierten Item-Faktor-Struktur entspricht und alle anhand der klassischen Kriterien ausgewählten Items enthält. Dieses Modell 1 weist bereits eine akzeptable Passung zu den Daten auf. Die Itembatterie konnte anhand der Diskriminationsparameter weiter verkleinert werden, das resultierende Modell 2 weist eine gute Passung zu den Daten auf. Es ist anzumerken, dass die standardisierten Diskriminationsparameter, die wie Trennschärfen interpretiert werden können, überwiegend hohe bis sehr hohe positive Werte aufweisen ($\bar{\lambda} = .71, SD = .10, \lambda_{\text{min}} = .50, \lambda_{\text{max}} = .84$). Ein plausibles einfaktorielles Alternativmodell, bei dem alle Items auf einen fiktiven Faktor laden, muss zugunsten des vierfaktoriellen Modells verworfen werden, Zufallsmodelle, bei denen die Items willkürlich vier Faktoren zugeordnet werden, wurden jedoch nicht gerechnet. Hier ist zu betonen, dass im Rahmen der konfirmatorischen Faktorenanalyse lediglich die Parameter der *spezifizierten* Item-Faktor-Beziehungen geschätzt werden: In der Ladungsmatrix sind andere als die spezifizierten Item-Faktor-Beziehungen auf null fixiert. Dadurch wird mit der Annahme des vierfaktoriellen Modells (Modell 2) gleichzeitig die Homogenität der Subskalen bestätigt, d. h. es ist davon auszugehen, dass keine Item-Multidimensionalitäten oder *cross-loadings* vorliegen. Die überwiegend hohen Faktorladungen sprechen für eine hohe konvergente Validität der Faktoren (Brown, 2006).

Weiterhin wurde der strukturelle Zusammenhang zwischen den latenten Faktoren untersucht. Es liegen nur in zwei Fällen signifikant von null verschiedene Korrelationen zwischen den Faktoren vor: Eine schwache Korrelation zwischen Expertenwissen und Daten als Evidenz ($r = .23$) sowie eine negative mittlere Korrelation zwischen den Kategorien Daten als Evidenz und Intuition ($r = -.50$). Dies ist theoriekonform: Ein Proband, der eher mit der Kategorie Intuition argumentiert, begründet eher weniger mit der Kategorie Daten als Evidenz. Allgemein kann das Fehlen bzw. die lediglich geringe bis mittlere Ausprägung der Korrelationen zwischen den latenten Faktoren als Hinweis für das Vorliegen diskriminanter Validität interpretiert werden (Brown, 2006, S. 131, nennt dafür als Kriterium $r < .8$). Alle vier Testteile messen voneinander unabhängige Konstrukte. Weiterhin wurde getestet, ob ein Modell mit unkorrelierten latenten Faktoren (Modell 4) eine bessere Passung zu den beobachteten Daten aufweist. Dieses Modell 4 muss jedoch aufgrund schlechterer Fitwerte abgelehnt werden. Innerhalb des Modells 2 wurden die Reliabilitäten der Testteile anhand der geschätzten Parameter bestimmt. Die Relia-

bilitäten liegen in einem Intervall von $.72 \leq \rho \leq .90$ und sind daher als hoch bis sehr hoch zu bezeichnen.

Bezüglich der Evaluation der Messinvarianz zeigt sich, dass sich alle vier Testteile über die Gruppen hinweg bis auf das Niveau skalarer Messinvarianz äquivalent verhalten. Es ist daher zum einen davon auszugehen, dass der Beantwortung des Tests in beiden Gruppen identische Konstrukte zugrunde liegen. Zum anderen aber sind die Messbeziehungen in Form der Zusammenhänge zwischen der latenten Variablen und dem Antwortverhalten auf manifester Ebene identisch. Mit dem Erreichen skalarer Messinvarianz ist eine notwendige Voraussetzung zum Vergleich von Mittelwertsunterschieden auf Gruppenebene in Folgestudien gegeben. Dieses Ergebnis ist insbesondere vor dem Hintergrund der Testkonzeption herauszustellen: Der vorliegende Test zur Erfassung der Verwendung bestimmter Argumente beim Experimentieren ist sowohl während der hier dokumentierten Entwicklung als auch in Folgestudien nur in Kombination mit einer *hypothesenprüfenden Experimentiersituation* einzusetzen. Eine Beeinflussung der zu messenden Konstrukte durch das Experiment selbst ist daher, wie eingangs geschildert, a priori nicht auszuschließen. In der vorliegenden Arbeit haben zwei Probandengruppen mit je einem Real- bzw. einem Computerexperiment gearbeitet. Die Invarianz der Messung in beiden Gruppen ist daher nicht nur ein Merkmal für den Test selbst, z. B. hinsichtlich des Itemfunktionierens, sondern gibt auch Aufschluss über die Konstrukte: Offenbar sind die Konstrukte in ihrer Struktur (nicht in ihrer Ausprägung!) von der vorhergehenden Experimentiersituation unbeeinflusst. Bei den Argumentkategorien scheint es sich offenbar um relativ stabile Konstrukte zu handeln, die strukturell nicht sensitiv auf äußere Gegebenheiten reagieren.

Insbesondere für die Testung der vierfaktoriellen Modelle muss jedoch die Modellgröße vor dem Hintergrund des Stichprobenumfangs diskutiert werden. Zur Stichprobenumfangsplanung im CFA-Kontext ist in der Literatur aktuell kein Konsens zu finden. (für einen Überblick zur Diskussion siehe Brown, 2006, S. 413). In der vorliegenden Arbeit wird jedoch insbesondere im Modell 1 eine hohe Zahl freier Parameter an einer verhältnismäßig kleinen Stichprobe geschätzt (df= 428, bei 99 freien Parameterm, siehe Tabelle 24). Dies kann zu einer unpräzisen Schätzung der Parameter führen. So kann es z. B. sein, dass Modell 1 rein aufgrund unzureichender Power sowie Unsicherheit in der Parameterschätzung abgelehnt werden musste, die Nichtpassung des Modells sollte daher nicht überbewertet werden. Auch muss berücksichtigt werden, dass bei dem hier geschilderten Vorgehen eine Anpassung des Testentwurfs an die Stichprobe erfolgt sein kann. Eine Replikation der Befunde in der Hauptstudie ist daher nötig.

Anhand der Berechnung psychometrischer Kenngrößen wurde die Itembatterie auf eine Größe von fünf Items pro Skala verkleinert. Es ist bei diesem Vorgehen davon auszugehen, dass es bis zu einem gewissen Grad zu einer inhaltlichen Einschränkung der Konstrukte gekommen ist, insbesondere weil die Itemselektion anhand von Kenngrößen erfolgte, die den Zusammenhang zwischen Item und Skala berücksichtigen. Es wurde dabei augenscheinlich geprüft, ob sich die inhaltliche Breite der erfassten Konstrukte nicht bedeutsam weiter verkleinert („construct underrepresentation", siehe American Educational Research Association, American Psychological Association, National Council on Measurement in Education und Joint Committee on Standards for Educational and Psychological Testing (2014, S. 10). Eine weitere Expertenstudie könnte diesen Sachverhalt evaluieren.

Aufgrund der vorhergehenden Argumentation ist daher abschließend davon auszugehen, dass es sich bei dem neu entwickelten Test zur Erfassung von Argumentationen beim hypothesenprüfenden Experimentieren um ein Instrument mit hoher inhaltlicher (aufgrund der Expertenstudie), faktorieller (aufgrund der Ergebnisse der CFA), diskriminanter (aufgrund der mittleren strukturellen Kovarianz) und konvergenter (aufgrund der hohen Faktorladungen) Validität sowie hoher Reliabilität handelt. Ein hohes Maß an Objektivität ist durch das fragenbogenbasierte Format gegeben. Die überprüften Konstrukte sind ferner invariant gegenüber der Testsituation (reales bzw. virtuelles Experiment), außerdem sind die Item-Faktor-Beziehungen bis auf das Niveau skalarer Invarianz äquivalent. Es muss daher konstatiert werden, dass der Test in hohem Maße für die Folgestudien geeignet ist.

B.8 FRAGEBÖGEN ZUR TESTENTWICKLUNG UND -EVALUATION

B.8.1 *Fragebogen zur Evaluation der inhaltlichen Validität*

HINWEIS Die hier abgedruckte Version des Fragebogens enthält eine Itemkodierung (z. B. `26-i6.08`), die vertikal unmittelbar vor den Itemtexten abgedruckt ist. Diese dient der Identifizierung der Items. In der Druckversion der Fragebögen zur Durchführung des Expertenratings war dieser Code ausgeblendet.

HUMBOLDT-UNIVERSITÄT ZU BERLIN
Lehrstuhl für Didaktik der Physik

Expertenrating der Testitems im Forschungsprojekt

„Argumentationsprozesse in Real- und Simulationsexperimenten“

Das vorliegende Testheft ist Teil eines aktuellen Forschungsprojekts am Lehrstuhl für Didaktik der Physik an der Humboldt-Universität zu Berlin. Mit dieser Studie soll die inhaltliche Validität der von uns entwickelten Items (Testfragen) untersucht werden. An dieser Studie werden Studierende des Lehramts Physik sowie wissenschaftliche Mitarbeiter aus den Didaktiken der Naturwissenschaften als „geschulte Experten“ teilnehmen. Dieses Testheft ist so aufgebaut, dass Sie an den jeweils notwendigen Stellen mit dem nötigen Wissen ausgestattet werden. Sie werden also zunächst immer erst zu „Experten“ ausgebildet und sollen **dann** die Items bezüglich eines Aspekts einschätzen.

Bitte gehen Sie nur linear vor (nicht blättern).

Vielen Dank für Ihre Unterstützung!

Prof. Burkhard Priemer, Jakob Bar und Tobias Ludwig

Personeninformationen

Damit wir die teilnehmende Expertengruppe einschätzen können, benötigen wir einige Daten von Ihnen:

_____
Name

_____
Alter

_____ _____
(Bisher) höchster Abschluss Fach bzw. Fächer

_____ _____
(Nächster) angestrebter Abschluss Fach bzw. Fächer

HUMBOLDT-UNIVERSITÄT ZU BERLIN
Lehrstuhl für Didaktik der Physik

Teil I: Hintergrundinformationen

Ziel und Ablauf des Forschungsprojekts
Ein Hauptziel der vorliegenden Studie ist es, zu untersuchen, ob Schülerinnen und Schüler, die mit realen und virtuellen Experimentierumgebungen selbstständig arbeiten, anhand der experimentellen Beobachtungen und Daten unterschiedlich argumentieren. Um eine Lerngelegenheit zu schaffen, die einen intensiven Argumentationsprozess anstößt, ist das verwendete Experiment so ausgelegt, dass die experimentellen Beobachtungen beim überwiegenden Teil der Probanden widersprüchlich zu bereits vorhandenen Präkonzepten bzw. Alltagsvorstellungen sind. Um diesen Konflikt stärker zu forcieren, werden die Probanden aufgefordert, im Vorfeld des Experiments, nach einer kurzen Einführung in den physikalischen Kontext, eine Hypothese aufzustellen. Nach dem Experiment werden die Probanden aufgefordert zu begründen, warum sie ihre eingangs aufgestellte Hypothese beibehalten oder ggf. verwerfen. Bisher wurden diese Argumentationen in Interviews erfasst. In Zukunft soll ein selbstauskunftsbasiertes Fragebogenverfahren eingesetzt werden. Die Items dieses Fragebogens sollen im vorliegenden Expertenrating im Hinblick auf inhaltliche Validität untersucht werden.

Begriffsdefinitionen

Begründungen: Die Gesamtheit aller Aussagen eines Probanden wird als Begründung bezeichnet. Eine Begründung kann sich dabei aus mehreren Argumentationen zusammensetzen.

Argumentationskategorie: Im vorliegenden Forschungsprojekt wurde ein System aus 10 Argumentationskategorien entwickelt und überprüft. Dieses System eignet sich dazu, die von Probanden gegebenen Aussagen zu erfassen und einzusortieren.

Teil II: Zuordnung der Items zu den Klassen

Klassen: Die Argumentationskategorien werden wiederum in zwei Klassen, eine **periphere** und eine **zentrale Klasse**, aufgeteilt:

Die **zentrale Klasse** umfasst diejenigen Argumentationskategorien, die auf eine rationale, d.h. vernunftgeleitete und sachliche Auseinandersetzung mit dem gesamten experimentellen Prozess (Durchführung des Experiments, Beobachtung und Sammeln der Daten sowie Evaluation der Daten) schließen lassen. Die zentrale Klasse zeichnet sich besonders durch eine kritische, intensive, aktive und motivierte Auseinandersetzung mit den Informationen aus. Argumentationen aus dieser Klasse stützen sich besonders auf Beobachtungen, Fakten und Evidenzen und sind daher meist intersubjektiv verständlich, d.h. für andere logisch nachvollziehbar und überprüfbar.

Die **periphere Klasse** umfasst hingegen alle Kategorien, die auf eine eher nichtrationale Auseinandersetzung mit dem experimentellen Prozess schließen lassen. Argumentationen in dieser Klasse orientieren sich oftmals an sog. peripheren Hinweisreizen (Cues), indem sie das Vertrauen in die Informationen und die Herkunft der Informationen analysieren und sich darauf beziehen. Argumentationen dieser Klasse können ferner auf eine gefühlsgeleitete, intuitive Auseinandersetzung mit dem experimentellen Prozess beruhen und sind allgemein geprägt von unkritischer und einer eher geringen rationalen Auseinandersetzung mit den Informationen.

Die folgende Tabelle stellt noch einmal die wesentlichen Schlüsselbegriffe der beiden Klassen dar.

Zentrale Klasse	Periphere Klasse
• rationale vernunftgeleitete sachliche kritische intensive aktive logische nachvollziehbare und intensive Auseinandersetzung mit dem Inhalt • Rückgriff auf Beobachtungen, Fakten und Evidenzen	• nicht-rational • intuitiv • gefühlsgeleitet • unkritisch • Eigenschaften des „Überbringers“ der Information wichtig, z.B. Sympathie und Expertise • Bewertung von oberflächlichen Merkmalen

AUFGABE:

Ordnen Sie nun auf Basis der Beschreibungen zur **peripheren** bzw. **zentralen Klasse** die folgenden Aussagen je einer Klasse zu!

#		Dieses Item gehört zu einer Argumentation aus der... ...peripheren Klasse.	...zentralen Klasse.	Dieses Item kann nicht eindeutig zugeordnet werden.
1-1.2.2	Wenn die Frage ist, ob ich den Daten vertrauen soll, entscheide ich aus dem Bauch heraus.	☐	☐	☐
2-1b.23	Die Messfehler waren so groß, dass sie bei meiner Entscheidung eine entscheidende Rolle gespielt haben.	☐	☐	☐
3-1b.04	Bei diesem Experiment gibt es Messungenauigkeiten, die ich bei meiner Schlussfolgerung berücksichtige.	☐	☐	☐
4-1.2.6	Richtig entschieden habe ich mich nicht. Ich habe nur geraten.	☐	☐	☐
5-1.2.0	Ich habe meine Vermutung beibehalten/verworfen. Ich habe im Experiment gemessen, dass das so stimmen müsste.	☐	☐	☐
6-1.2.1	Bei der Entscheidung habe ich mich ganz auf mein Gefühl verlassen.	☐	☐	☐
7-1b.01	Die Messdaten sind so eindeutig, dass ich das Beibehalten/Verwerfen meiner Vermutung damit stützen kann.	☐	☐	☐
8-1.1.9	Ich wechsle/beibehalte meine Vermutung wegen der Messwerte auf der Stoppuhr.	☐	☐	☐
9-1.1.25	Ich habe eher geraten als richtig über die experimentellen Beobachtungen nachgedacht.	☐	☐	☐
10-1b.07	Ich bin unsicher, ob ich meine Vermutung wechsle/beibehalte, weil die Messwerte leichte Abweichungen haben.	☐	☐	☐
11-1b.18	Meine Entscheidung, die Vermutung zu verwerfen/beizubehalten, wird beeinflusst durch die Ungenauigkeiten des Experiments.	☐	☐	☐
12-dummy4	Bevor ich die Entscheidung treffen kann, muss mir noch einmal jemand sagen, dass ich richtig gearbeitet habe.	☐	☐	☐
13-1.1.3	Beim Treffen der Entscheidung folge ich meinem Gefühl.	☐	☐	☐

#		Dieses Item gehört zu einer Argumentation aus der... ...peripheren Klasse.	...zentralen Klasse.	Dieses Item kann nicht eindeutig zugeordnet werden.
14 t4.03	Anhand meiner Messwerte sehe ich, dass es richtig ist, meine Vermutung beizubehalten/zu verwerfen.	☐	☐	☐
15 t6.09	Bei meiner Entscheidung berücksichtige ich ganz besonders die deutlichen Messwerte.	☐	☐	☐
16 t1.20	Ich ziehe die Schlussfolgerungen aus dem Experiment eher aufgrund meiner Gefühle.	☐	☐	☐
17 dummy1	Ich habe mich bei dieser Entscheidung auf mein Bauchgefühl und die experimentellen Beobachtungen gestützt.	☐	☐	☐
18 t1.10	Ich kann nicht sagen, warum ich mich für das Beibehalten/Verwerfen meiner Vermutung entschieden habe.	☐	☐	☐
19 t7b.17	Bei meiner Entscheidung berücksichtige ich auch, dass das Experiment ungenau ist.	☐	☐	☐
20 t6.17	Ich wechsle/beibehalte meine Vermutung, weil ich das durch die gemessenen Zeiten begründen kann.	☐	☐	☐
21 t7b.12	Die Entscheidung, meine Vermutung zu verwerfen/beizubehalten fiel mir schwer, weil die Ergebnisse nicht genau waren.	☐	☐	☐
22 t8.21	Wegen dem, was ich beim Messen herausgefunden habe, wechsle/verwerfe ich meine Vermutung.	☐	☐	☐
23 t7.04	Ohne großes Nachdenken erscheint es mir in diesem Fall richtig, meine Vermutung beizubehalten/zu verwerfen.	☐	☐	☐
24 t6.15	Ich vertraue bei meiner Entscheidung hier nur auf die Beobachtungen, die ich gemacht habe.	☐	☐	☐
25 t8.26	Wenn ich die Messwerte betrachte, sind diese alleine ausreichend, um meine Entscheidung zum Beibehalten/Verwerfen zu begründen.	☐	☐	☐
26 t6.08	Bei meiner Entscheidung berücksichtige ich, dass das Experiment bereits von Wissenschaftlern erdacht und überprüft wurde.	☐	☐	☐

#		Dieses Item gehört zu einer Argumentation aus der... ...peripheren Klasse.	...zentralen Klasse.	Dieses Item kann nicht eindeutig zugeordnet werden.
27 t8.22	Ich denke erst über die Messdaten nach, bevor ich eine Entscheidung treffe.	☐	☐	☐
28 t7b.19	Das Experiment ist sehr ungenau, das habe ich jedoch bei meiner Entscheidung nicht berücksichtigt.	☐	☐	☐
29 t8.15	Bei der Entscheidung berücksichtige ich, dass das Experiment ja schon einmal von anderen überprüft wurde.	☐	☐	☐
30 dummy3	Ich habe die Daten aus guten Gründen komplett ignoriert.	☐	☐	☐
31 t8.14	Die Messdaten spielen bei meiner Entscheidung die größte Rolle.	☐	☐	☐
32 t8.18	Ich berücksichtige bei meiner Entscheidung, dass die Menschen, die das Experiment aufgebaut haben, mehr Fachwissen haben als ich.	☐	☐	☐
33 t8.12	Das Experiment ist ja schon einmal geprüft worden, deshalb vertraue ich darauf, dass es fehlerfrei ist. Dies berücksichtige ich bei meiner Entscheidung.	☐	☐	☐
34 t4.13	Diese Ergebnisse haben mir gezeigt, dass es richtig ist, meine Vermutung beizubehalten/zuverwerfen.	☐	☐	☐
35 t7b.20	Beim Messen beobachtete ich Ungenauigkeiten. Das berücksichtige ich in hohem Maße bei meiner Entscheidung.	☐	☐	☐
36 t7b.10	Bei meiner Entscheidung zum Beibehalten/Verwerfen meiner Vermutung berücksichtige ich, dass die Messwerte deutlich schwanken.	☐	☐	☐
37 t6.07	An meinen Messwerten sehe ich, ob meine Vermutung richtig/falsch war.	☐	☐	☐
38 t1.16	Bei meiner Entscheidung spielen Gefühle eine große Rolle.	☐	☐	☐
39 t8.10	Ich berücksichtige bei meiner Entscheidung, dass man dem Experiment vertrauen kann.	☐	☐	☐

#		Dieses Item gehört zu einer Argumentation aus der... ...peripheren Klasse.	...zentralen Klasse.	Dieses Item kann nicht eindeutig zugeordnet werden.
v40.02	Ich habe die Schwingungsdauer gemessen und daher muss ich meine Vermutung verwerfen/beibehalten.	☐	☐	☐
v1.11	Um für mich diese Entscheidung zum Beibehalten/Verwerfen meiner Vermutung zu begründen, habe ich besonders intensiv über die Messdaten nachgedacht.	☐	☐	☐
v2.05	Meine Messdaten sind der Grund, warum ich meine Vermutung verwerfe/beibehalte.	☐	☐	☐
v3.11	Bei meiner Entscheidung berücksichtige ich, dass bestimmt keine Fehler im Aufbau des Experiments enthalten sind.	☐	☐	☐
v4.07	Bei meiner Entscheidung zum Beibehalten/Verwerfen meiner Vermutung berücksichtige ich, dass das Experiment an einer Universität entwickelt wurde.	☐	☐	☐
v5.22	Ich habe die Messungenauigkeiten beim Experimentieren in hohem Maße berücksichtigt.	☐	☐	☐
v6.13	Es gibt zwar immer Messfehler - in diesem Fall haben Sie aber meine Entscheidung nicht beinflusst.	☐	☐	☐
v7.12	Bei der Entscheidung zum Beibehalten/Verwerfen meiner Vermutung habe ich deutlich auf meine Gefühle gehört.	☐	☐	☐
v8.04	Die Messwerte sind ausreichend um meine Vermutung zu wechseln/beizubehalten.	☐	☐	☐
v9.19	Ich habe mir über das Experiment sehr schnell einen Eindruck gebildet und diesen bei meiner Entscheidung berücksichtigt.	☐	☐	☐
v10.14	Bei dieser Entscheidung ist es sinnvoll, sich auf sein Gefühl zu verlassen.	☐	☐	☐
v11.08	Die Ungenauigkeiten beim Experimentieren erschweren meine Entscheidung.	☐	☐	☐
v12.05	Bei meiner Entscheidung berücksichtige ich, dass das Experiment bestimmt richtig funktioniert.	☐	☐	☐

#		Dieses Item gehört zu einer Argumentation aus der... ...peripheren Klasse.	...zentralen Klasse.	Dieses Item kann nicht eindeutig zugeordnet werden.
v33.17	Ich habe mich spontan und schnell entschieden, ob ich meine Vermutung wechsle/beibehalte.	☐	☐	☐
v44.24	Ich stütze meine Entscheidung zum Beibehalten/Wechseln meiner Vermutung auf die Fakten, die das Experiment mir liefert.	☐	☐	☐
v55.04	Beim Experimentieren bin ich davon ausgegangen, dass das Experiment richtige Ergebnisse liefert.	☐	☐	☐
v66.09	Meine Entscheidung, die Vermutung beizubehalten/zu verwerfen ist eine Reaktion auf die Messdaten.	☐	☐	☐
v77.16	Ich vertraue auf den Entwickler des Experiments. Das beeinflusst meine Entscheidung, die Vermutung zu verwerfen/beizubehalten.	☐	☐	☐
v88.08	Das Ergebnisse zeigen mir eindeutig, dass ich meine Vermutung verwerfen/beibehalten muss.	☐	☐	☐
v99.05	Ich denke gar nicht nach, wenn ich mich für eine Vermutung entscheide.	☐	☐	☐
v101.07	Ich entscheide mich ohne großes Nachdenken, wenn ich aus den Daten Schlüsse ziehe.	☐	☐	☐
v111.15	Bei dieser Entscheidung zum Beibehalten/Verwerfen habe ich aus dem Bauch heraus entschieden.	☐	☐	☐
v122.16	Ich berücksichtige, dass das Experiment von Wissenschaftlern mitgebracht wurde, die an der Humboldt-Universität arbeiten.	☐	☐	☐
v133.09	So ein Experiment ist ja meistens richtig. Dies berücksichtige ich bei meiner Entscheidung.	☐	☐	☐
v144.03	Ich traue dem Experiment nicht ganz, weil es von einem Menschen konstruiert wurde, der sich auch irren kann.	☐	☐	☐
v155.02	Bei diesem Experiment ist es wichtig, die Messung mehrfach zu wiederholen, um sichere Aussagen machen zu können.	☐	☐	☐

#		Dieses Item gehört zu einer Argumentation aus der... ...peripheren Klasse.	...zentralen Klasse.	Dieses Item kann nicht eindeutig zugeordnet werden.
V66-fr18	Die Auswertung meiner Messung ist der Grund für meine Entscheidung zum Beibehalten/Wechseln meiner Vermutung.	☐	☐	☐
V72-fr03	Ich höre stark auf mein Bauchgefühl, wenn ich aus dem Experiment Schlüsse ziehe.	☐	☐	☐
V64-fr09	Ich habe die schwankenden Messwerte bei meiner Entscheidung zum Beibehalten/Verwerfen meiner Vermutung nicht berücksichtigt.	☐	☐	☐
V69-fr06	Wenn ich die Messdaten interpretiere, muss ich auch berücksichtigen, dass das Experiment etwas ungenau ist.	☐	☐	☐
V70-fr05	Ich berücksichtige bei meiner Entscheidung, dass das Messen mit der Stoppuhr nicht ganz genau ist.	☐	☐	☐
V71-fr23	Wenn es darum geht, ob ich dem Experiment vertrauen soll, entscheide ich aus dem Bauch heraus.	☐	☐	☐
V73-fr09	Bei der Entscheidung berücksichtige ich stark mein Gefühl.	☐	☐	☐
V75-fr06	Bei meiner Entscheidung berücksichtige ich, dass das Experiment professionell aussah.	☐	☐	☐
V74-fr10	Bevor ich die Entscheidung zum Beibehalten/Verwerfen meiner Vermutung treffe, denke ich zunächst einmal gründlich über die Messdaten nach.	☐	☐	☐
V75-summ2	Ich glaube, dass ich nicht gut experimentieren kann, das berücksichtige ich bei meiner Entscheidung.	☐	☐	☐
V76-fr06	Die Gründe für meine Entscheidung zum Beibehalten/Verwerfen meiner Vermutung sind mir nicht richtig bewusst.	☐	☐	☐
V77-fr02	Mein Gefühl sagt mir, dass ich meine Vermutung in diesem Fall beibehalten/wechseln sollte.	☐	☐	☐
V78-fr22	Anhand der Messwerte überprüfe ich, ob ich meine Vermutung wechsle oder beibehalten soll.	☐	☐	☐

#		Dieses Item gehört zu einer Argumentation aus der... ...peripheren Klasse.	...zentralen Klasse.	Dieses Item kann nicht eindeutig zugeordnet werden.
V79-fr18	Ich habe mich gefühlsmäßig für die Antwort entschieden, die mir am meisten zusagt.	☐	☐	☐
V80-fr24	Bei meiner Entscheidung habe ich gar nicht richtig nachgedacht, ich habe eher geraten.	☐	☐	☐
V81-fr07	Bei meiner Entscheidung berücksichtige ich nicht, dass hier schon eine Menge Wissen im Experiment enthalten ist.	☐	☐	☐
V82-fr08	Ich habe ein Gespür dafür, dass meine Vermutung richtig/falsch ist	☐	☐	☐
V83-fr11	Kleine Abstände der Messwerte verunsichern mich bei meiner Entscheidung, ob ich meine Vermutung beibehalte oder verwerfe.	☐	☐	☐
V84-fr15	Es gibt hier bestimmt auch Messfehler. Das berücksichtige ich beim Beibehalten/Verwerfen meiner Entscheidung.	☐	☐	☐
V85-fr11	Bei der Entscheidung zum Beibehalten/Verwerfen meiner Vermutung verlasse ich mich auf mein Gespür.	☐	☐	☐
V86-fr14	Ich berücksichtige bei meiner Entscheidung, dass das Messen mit der Stoppuhr nicht ganz genau ist.	☐	☐	☐
V87-fr21	Bei meiner Entscheidung habe ich nicht beachtet, dass das Experiment ungenau ist.	☐	☐	☐
V88-fr16	Bei der Entscheidung zum Wechseln/Beibehalten meiner Vermutung berücksichtige ich, dass man immer Ungenauigkeiten beim Experimentieren hat.	☐	☐	☐
V89-fr25	Ich entscheide mich für/gegen meine Vermutung nur anhand meiner experimentellen Beobachtungen.	☐	☐	☐
V90-fr23	Ich habe beim Stoppen Zeiten gemessen und diese Messwerte veranlassen mich, meine Vermutung beizubehalten/zu wechseln.	☐	☐	☐
V91-fr14	Bei meiner Entscheidung berücksichtige ich, dass das Experiment ja bestimmt nicht falsch sein wird.	☐	☐	☐

#		Dieses Item gehört zu einer Argumentation aus der... ...peripheren Klasse.	...zentralen Klasse.	Dieses Item kann nicht eindeutig zugeordnet werden.
2018.11	Ich wechsle/behalte meine Vermutung wegen des Experimentiergebnisses.	☐	☐	☐

Teil III: Zuordnung der Items zu den Argumentationskategorien

Im vorliegenden Forschungsprojekt werden vier von zehn möglichen Argumentationskategorienategorien näher untersucht. Es handelt sich dabei um die Kategorien **Evidenz, Expertenwissen, Intuition** und **Erwähnung der Relevanz von Messunsicherheiten**. Sie sollen im Folgenden die Items, die sie bereits zu den Klassen zugeordnet haben nun zu den Argumentationskategorien zuordnen. Dazu werden zunächst die Argumentationskategorien beschrieben.

Kategorie 1: Intuition
Die Kategorie „Intuition" umfasst alle Aussagen, die darauf schließen lassen, dass die Entscheidung zum Wechseln bzw. Beibehalten einer Hypothese ohne diskursiven Gebrauch des Verstandes sondern durch eher gefühlsmäßige Eingebungen und Ahnungen zustande kommt. Die Entscheidung wird ohne expliziten Rückgriff auf die experimentellen Beobachtungen bzw. Daten getroffen. Argumentativ berufen sich Aussagen dieser Kategorie oft auf das ‚Bauchgefühl'. Auch ‚unbewusste' und nicht benannte bzw. benennbare Aspekte führen zu einer intuitiven Argumentation.

Kategorie 6: Expertenwissen
Die Kategorie „Expertenwissen" umfasst Aussagen, die das Wissen eines ausgewiesenen Experten (im Experiment) berücksichtigen. Im Falle der Experiment-Simulation meint dies eine Bezugnahme auf das zugrundeliegende analytische Modell der physikalischen Realität und berücksichtigt die Tatsache, dass dieses bereits durch eine Person mit größerem physikalischen Verständnis programmiert wurde. Im Falle des Realexperiments umfasst diese Kategorie Aussagen, die z.B. das Vertrauen in den experimentellen Aufbau in Bezug auf die Herkunft des Experiments (z.B. Wissenschaftler an einer Universität) bewerten.

Kategorie 7b: Erwähnung der Relevanz von Messunsicherheiten
Die Kategorie „Erwähnung der Relevanz von Messunsicherheiten" umfasst Aussagen, die darauf schließen lassen, dass eine Auseinandersetzung im Hinblick auf die Verlässlichkeit der Interpretation der Messdaten stattfindet. Diese kritische Herangehensweise kann sich auf schwankende Messwerte bei der Wiederholung des Experiments oder auf die Genauigkeit des Messprozesses beziehen. Diese Kategorie beinhaltet *keine* Aussagen, die sich aus einer Beurteilung der eigenen experimentellen Fähigkeiten ergeben und sich damit auf personenbezogene Unsicherheitsquellen beziehen.

Kategorie 8: Evidenz
Die Argumentationskategorie „Evidenz" umfasst alle Aussagen, die darauf schließen lassen, dass die Entscheidung zum Beibehalten oder Wechseln der Hypothese auf Evidenzen beruhen. Im Allgemeinen werden „Evidenzen" als empirische Befunde verstanden, die (wissenschaftliche) Erkenntnis rechtfertigen. Im vorliegenden Kontext des selbstständigen Experimentierens wird daher z.B. konkret auf experimentell ermittelte Daten bzw. Messwerte Bezug genommen. Es ist zu beachten, dass diese Kategorie sowohl Aussagen umfasst, die konkret auf gemessene Werte Bezug nehmen, als auch Aussagen, welche die gemessenen Daten nur implizit berücksichtigen („Es wurde ja durch das Experiment bewiesen, dass die Zeit immer gleich war".)

Die folgende Tabelle fasst noch einmal die wesentliche Merkmale der Kategoriebeschreibungen zusammen.

Intuition	Expertenwissen	Messunsicherheiten	Evidenz
• gefühlsmäßige Eingebungen • Ahnungen • ohne Rückgriff auf experimentelle Beobachtungen • Bauchgefühl • unbewusste Aspekte • nicht benannte oder benennbare Aspekte • Raten • schnelle Entscheidungen	• Entscheidungen aufgrund im Experiment enthaltener ausgewiesener Expertise • Analyse der Herkunft des Experiments	• Beachtung von Messunsicherheiten • kritische Analyse der Messdaten • Bezugnahme auf Genauigkeit des Messprozesses	• Entscheidungen beruhen auf Beobachtungen, Daten, Evidenzen. • implizite und explizite Bezugnahme auf Evidenzen

AUFGABE:

Ordnen Sie nun auf Basis der Kategoriebeschreibungen die folgenden Aussagen je einer Kategorie zu.

#		Dieses Item gehört zu einer Argumentation aus der Kategorie... Intuition	Expertenwissen	Messunsicherheit	Evidenz	Dieses Item kann nicht eindeutig zugeordnet werden.
1	Wenn die Frage ist, ob ich den Daten vertrauen soll, entscheide ich aus dem Bauch heraus.	☐	☐	☐	☐	☐
2	Die Messfehler waren so groß, dass sie bei meiner Entscheidung eine entscheidende Rolle gespielt haben.	☐	☐	☐	☐	☐
3	Bei diesem Experiment gibt es Messungenauigkeiten, die ich bei meiner Schlussfolgerung berücksichtige.	☐	☐	☐	☐	☐
4	Richtig entschieden habe ich mich nicht. Ich habe nur geraten.	☐	☐	☐	☐	☐
5	Ich habe meine Vermutung beibehalten/verworfen. Ich habe im Experiment gemessen, dass das so stimmen müsste.	☐	☐	☐	☐	☐
6	Bei der Entscheidung habe ich mich ganz auf mein Gefühl verlassen.	☐	☐	☐	☐	☐
7	Die Messdaten sind so eindeutig, dass ich das Beibehalten/Verwerfen meiner Vermutung damit stützen kann.	☐	☐	☐	☐	☐
8	Ich wechsle/beibehalte meine Vermutung wegen der Messwerte auf der Stoppuhr.	☐	☐	☐	☐	☐
9	Ich habe eher geraten als richtig über die experimentellen Beobachtungen nachgedacht.	☐	☐	☐	☐	☐
10	Ich bin unsicher, ob ich meine Vermutung wechsle/beibehalte, weil die Messwerte leichte Abweichungen haben.	☐	☐	☐	☐	☐
11	Meine Entscheidung, die Vermutung zu verwerfen/beizubehalten, wird beeinflusst durch die Ungenauigkeiten des Experiments.	☐	☐	☐	☐	☐
12	Bevor ich die Entscheidung treffen kann, muss mir noch einmal jemand sagen, dass ich richtig gearbeitet habe.	☐	☐	☐	☐	☐

#		Dieses Item gehört zu einer Argumentation aus der Kategorie...				Dieses Item kann nicht eindeutig zugeordnet werden.
		Intuition	Expertenwissen	Messunsicherheit	Evidenz	
13 (v1.13)	Beim Treffen der Entscheidung folge ich meinem Gefühl.	☐	☐	☐	☐	☐
14 (v8.03)	Anhand meiner Messwerte sehe ich, dass es richtig ist, meine Vermutung beizubehalten/zu verwerfen.	☐	☐	☐	☐	☐
15 (v8.06)	Bei meiner Entscheidung berücksichtige ich ganz besonders die deutlichen Messwerte.	☐	☐	☐	☐	☐
16 (v1.20)	Ich ziehe die Schlussfolgerungen aus dem Experiment eher aufgrund meiner Gefühle.	☐	☐	☐	☐	☐
17 (dummy)	Ich habe mich bei dieser Entscheidung auf mein Bauchgefühl und die experimentellen Beobachtungen gestützt.	☐	☐	☐	☐	☐
18 (v1.10)	Ich kann nicht sagen, warum ich mich für das Beibehalten/Verwerfen meiner Vermutung entschieden habe.	☐	☐	☐	☐	☐
19 (v7b.17)	Bei meiner Entscheidung berücksichtige ich auch, dass das Experiment ungenau ist.	☐	☐	☐	☐	☐
20 (v8.17)	Ich wechsle/beibehalte meine Vermutung, weil ich das durch die gemessenen Zeiten begründen kann.	☐	☐	☐	☐	☐
21 (v7b.12)	Die Entscheidung, meine Vermutung zu verwerfen/beizubehalten fiel mir schwer, weil die Ergebnisse nicht genau waren.	☐	☐	☐	☐	☐
22 (v8.21)	Wegen dem, was ich beim Messen herausgefunden habe, wechsle/verwerfe ich meine Vermutung.	☐	☐	☐	☐	☐
23 (v1.04)	Ohne großes Nachdenken erscheint es mir in diesem Fall richtig, meine Vermutung beizubehalten/zu verwerfen.	☐	☐	☐	☐	☐
24 (v8.15)	Ich vertraue bei meiner Entscheidung hier nur auf die Beobachtungen, die ich gemacht habe.	☐	☐	☐	☐	☐

#		Dieses Item gehört zu einer Argumentation aus der Kategorie...				Dieses Item kann nicht eindeutig zugeordnet werden.
		Intuition	Expertenwissen	Messunsicherheit	Evidenz	
25 (v9.26)	Wenn ich die Messwerte betrachte, sind diese alleine ausreichend, um meine Entscheidung zum Beibehalten/Verwerfen zu begründen.	☐	☐	☐	☐	☐
26 (v8.08)	Bei meiner Entscheidung berücksichtige ich, dass das Experiment bereits von Wissenschaftlern erdacht und überprüft wurde.	☐	☐	☐	☐	☐
27 (v8.12)	Ich denke erst über die Messdaten nach, bevor ich eine Entscheidung treffe.	☐	☐	☐	☐	☐
28 (v7b.19)	Das Experiment ist sehr ungenau, das habe ich jedoch bei meiner Entscheidung nicht berücksichtigt.	☐	☐	☐	☐	☐
29 (v9.13)	Bei der Entscheidung berücksichtige ich, dass das Experiment ja schon einmal von anderen überprüft wurde.	☐	☐	☐	☐	☐
30 (dummy)	Ich habe die Daten aus guten Gründen komplett ignoriert.	☐	☐	☐	☐	☐
31 (v8.14)	Die Messdaten spielen bei meiner Entscheidung die größte Rolle.	☐	☐	☐	☐	☐
32 (v9.18)	Ich berücksichtige bei meiner Entscheidung, dass die Menschen, die das Experiment aufgebaut haben, mehr Fachwissen haben als ich.	☐	☐	☐	☐	☐
33 (v9.12)	Das Experiment ist ja schon einmal geprüft worden, deshalb vertraue ich darauf, dass es fehlerfrei ist. Dies berücksichtige ich bei meiner Entscheidung.	☐	☐	☐	☐	☐
34 (v8.13)	Diese Ergebnisse haben mir gezeigt, dass es richtig ist, meine Vermutung beizubehalten/zuverwerfen.	☐	☐	☐	☐	☐
35 (v7b.20)	Beim Messen beobachtete ich Ungenauigkeiten. Das berücksichtige ich in hohem Maße bei meiner Entscheidung.	☐	☐	☐	☐	☐
36 (v7b.10)	Bei meiner Entscheidung zum Beibehalten/Verwerfen meiner Vermutung berücksichtige ich, dass die Messwerte deutlich schwanken.	☐	☐	☐	☐	☐

#		Dieses Item gehört zu einer Argumentation aus der Kategorie...				Dieses Item kann nicht eindeutig zugeordnet werden.
		Intuition	Expertenwissen	Messunsicherheit	Evidenz	
B7 A0.07	An meinen Messwerten sehe ich, ob meine Vermutung richtig/falsch war.	☐	☐	☐	☐	☐
B8 A1.16	Bei meiner Entscheidung spielen Gefühle eine große Rolle.	☐	☐	☐	☐	☐
B9 A0.10	Ich berücksichtige bei meiner Entscheidung, dass man dem Experiment vertrauen kann.	☐	☐	☐	☐	☐
B10 A1.02	Ich habe die Schwingungsdauer gemessen und daher muss ich meine Vermutung verwerfen/beibehalten.	☐	☐	☐	☐	☐
B11 A1.11	Um für mich diese Entscheidung zum Beibehalten/Verwerfen meiner Vermutung zu begründen, habe ich besonders intensiv über die Messdaten nachgedacht.	☐	☐	☐	☐	☐
B12 A1.05	Meine Messdaten sind der Grund, warum ich meine Vermutung verwerfe/beibehalte.	☐	☐	☐	☐	☐
B13 A0.11	Bei meiner Entscheidung berücksichtige ich, dass bestimmt keine Fehler im Aufbau des Experiments enthalten sind.	☐	☐	☐	☐	☐
B14 A0.07	Bei meiner Entscheidung zum Beibehalten/Verwerfen meiner Vermutung berücksichtige ich, dass das Experiment an einer Universität entwickelt wurde.	☐	☐	☐	☐	☐
B15 A1.7b.22	Ich habe die Messungenauigkeiten beim Experimentieren in hohem Maße berücksichtigt.	☐	☐	☐	☐	☐
B16 A1.7b.13	Es gibt zwar immer Messfehler - in diesem Fall haben Sie aber meine Entscheidung nicht beinflusst.	☐	☐	☐	☐	☐
B17 A1.12	Bei der Entscheidung zum Beibehalten/Verwerfen meiner Vermutung habe ich deutlich auf meine Gefühle gehört.	☐	☐	☐	☐	☐
B18 A0.04	Die Messwerte sind ausreichend um meine Vermutung zu wechseln/beizubehalten.	☐	☐	☐	☐	☐

#		Dieses Item gehört zu einer Argumentation aus der Kategorie...				Dieses Item kann nicht eindeutig zugeordnet werden.
		Intuition	Expertenwissen	Messunsicherheit	Evidenz	
B19 A1.19	Ich habe mir über das Experiment sehr schnell einen Eindruck gebildet und diesen bei meiner Entscheidung berücksichtigt.	☐	☐	☐	☐	☐
B20 A1.14	Bei dieser Entscheidung ist es sinnvoll, sich auf sein Gefühl zu verlassen.	☐	☐	☐	☐	☐
B21 A1.7b.38	Die Ungenauigkeiten beim Experimentieren erschweren meine Entscheidung.	☐	☐	☐	☐	☐
B22 A1.05	Bei meiner Entscheidung berücksichtige ich, dass das Experiment bestimmt richtig funktioniert.	☐	☐	☐	☐	☐
B23 A1.17	Ich habe mich spontan und schnell entschieden, ob ich meine Vermutung wechsle/beibehalte.	☐	☐	☐	☐	☐
B24 A1.24	Ich stütze meine Entscheidung zum Beibehalten/Wechseln meiner Vermutung auf die Fakten, die das Experiment mir liefert.	☐	☐	☐	☐	☐
B25 A1.04	Beim Experimentieren bin ich davon ausgegangen, dass das Experiment richtige Ergebnisse liefert.	☐	☐	☐	☐	☐
B26 A1.09	Meine Entscheidung, die Vermutung beizubehalten/zu verwerfen ist eine Reaktion auf die Messdaten.	☐	☐	☐	☐	☐
B27 A1.16	Ich vertraue auf den Entwickler des Experiments. Das beeinflusst meine Entscheidung, die Vermutung zu verwerfen/beizubehalten.	☐	☐	☐	☐	☐
B28 A1.08	Das Ergebnisse zeigen mir eindeutig, dass ich meine Vermutung verwerfen/beibehalten muss.	☐	☐	☐	☐	☐
B29 A1.06	Ich denke gar nicht nach, wenn ich mich für eine Vermutung entscheide.	☐	☐	☐	☐	☐
B30 A1.07	Ich entscheide mich ohne großes Nachdenken, wenn ich aus den Daten Schlüsse ziehe.	☐	☐	☐	☐	☐

#		Dieses Item gehört zu einer Argumentation aus der Kategorie...				Dieses Item kann nicht eindeutig zugeordnet werden.
		Intuition	Expertenwissen	Messunsicherheit	Evidenz	
I01.1.15	Bei dieser Entscheidung zum Beibehalten/Verwerfen habe ich aus dem Bauch heraus entschieden.	☐	☐	☐	☐	☐
I02.0.15	Ich berücksichtige, dass das Experiment von Wissenschaftlern mitgebracht wurde, die an der Humboldt-Universität arbeiten.	☐	☐	☐	☐	☐
I03.4.09	So ein Experiment ist ja meistens richtig. Dies berücksichtige ich bei meiner Entscheidung.	☐	☐	☐	☐	☐
I04.4.03	Ich traue dem Experiment nicht ganz, weil es von einem Menschen konstruiert wurde, der sich auch irren kann.	☐	☐	☐	☐	☐
I05.7b.02	Bei diesem Experiment ist es wichtig, die Messung mehrfach zu wiederholen, um sichere Aussagen machen zu können.	☐	☐	☐	☐	☐
I06.4.16	Die Auswertung meiner Messung ist der Grund für meine Entscheidung zum Beibehalten/Wechseln meiner Vermutung.	☐	☐	☐	☐	☐
I07.1.03	Ich höre stark auf mein Bauchgefühl, wenn ich aus dem Experiment Schlüsse ziehe.	☐	☐	☐	☐	☐
I08.7b.09	Ich habe die schwankenden Messwerte bei meiner Entscheidung zum Beibehalten/Verwerfen meiner Vermutung nicht berücksichtigt.	☐	☐	☐	☐	☐
I09.7b.06	Wenn ich die Messdaten interpretiere, muss ich auch berücksichtigen, dass das Experiment etwas ungenau ist.	☐	☐	☐	☐	☐
I10.7b.05	Ich berücksichtige bei meiner Entscheidung, dass das Messen mit der Stoppuhr nicht ganz genau ist.	☐	☐	☐	☐	☐
I11.1.23	Wenn es darum geht, ob ich dem Experiment vertrauen soll, entscheide ich aus dem Bauch heraus.	☐	☐	☐	☐	☐
I12.1.09	Bei der Entscheidung berücksichtige ich stark mein Gefühl.	☐	☐	☐	☐	☐

#		Dieses Item gehört zu einer Argumentation aus der Kategorie...				Dieses Item kann nicht eindeutig zugeordnet werden.
		Intuition	Expertenwissen	Messunsicherheit	Evidenz	
I13.0.06	Bei meiner Entscheidung berücksichtige ich, dass das Experiment professionell aussah.	☐	☐	☐	☐	☐
I14.4.10	Bevor ich die Entscheidung zum Beibehalten/Verwerfen meiner Vermutung treffe, denke ich zunächst einmal gründlich über die Messdaten nach.	☐	☐	☐	☐	☐
I15.dummy2	Ich glaube, dass ich nicht gut experimentieren kann, das berücksichtige ich bei meiner Entscheidung.	☐	☐	☐	☐	☐
I16.1.06	Die Gründe für meine Entscheidung zum Beibehalten/Verwerfen meiner Vermutung sind mir nicht richtig bewusst.	☐	☐	☐	☐	☐
I17.1.02	Mein Gefühl sagt mir, dass ich meine Vermutung in diesem Fall beibehalten/wechseln sollte.	☐	☐	☐	☐	☐
I18.4.22	Anhand der Messwerte überprüfe ich, ob ich meine Vermutung wechsle oder beibehalten soll.	☐	☐	☐	☐	☐
I19.1.18	Ich habe mich gefühlsmäßig für die Antwort entschieden, die mir am meisten zusagt.	☐	☐	☐	☐	☐
I20.1.24	Bei meiner Entscheidung habe ich gar nicht richtig nachgedacht, ich habe eher geraten.	☐	☐	☐	☐	☐
I21.4.17	Bei meiner Entscheidung berücksichtige ich nicht, dass hier schon eine Menge Wissen im Experiment enthalten ist.	☐	☐	☐	☐	☐
I22.1.08	Ich habe ein Gespür dafür, dass meine Vermutung richtig/falsch ist	☐	☐	☐	☐	☐
I23.7b.11	Kleine Abstände der Messwerte verunsichern mich bei meiner Entscheidung, ob ich meine Vermutung beibehalte oder verwerfe.	☐	☐	☐	☐	☐
I24.7b.15	Es gibt hier bestimmt auch Messfehler. Das berücksichtige ich beim Beibehalten/Verwerfen meiner Entscheidung.	☐	☐	☐	☐	☐

HUMBOLDT-UNIVERSITÄT ZU BERLIN

Lehrstuhl für Didaktik der Physik

#		Dieses Item gehört zu einer Argumentation aus der Kategorie...				Dieses Item kann nicht eindeutig zugeordnet werden.
		Intuition	Expertenwissen	Messunsicherheit	Evidenz	
16	Bei der Entscheidung zum Beibehalten/Verwerfen meiner Vermutung verlasse ich mich auf mein Gespür.	☐	☐	☐	☐	☐
14	Ich berücksichtige bei meiner Entscheidung, dass das Messen mit der Stoppuhr nicht ganz genau ist.	☐	☐	☐	☐	☐
21	Bei meiner Entscheidung habe ich nicht beachtet, dass das Experiment ungenau ist.	☐	☐	☐	☐	☐
18	Bei der Entscheidung zum Wechseln/Beibehalten meiner Vermutung berücksichtige ich, dass man immer Ungenauigkeiten beim Experimentieren hat.	☐	☐	☐	☐	☐
25	Ich entscheide mich für/gegen meine Vermutung nur anhand meiner experimentellen Beobachtungen.	☐	☐	☐	☐	☐
23	Ich habe beim Stoppen Zeiten gemessen und diese Messwerte veranlassen mich, meine Vermutung beizubehalten/zu wechseln.	☐	☐	☐	☐	☐
14	Bei meiner Entscheidung berücksichtige ich, dass das Experiment ja bestimmt nicht falsch sein wird.	☐	☐	☐	☐	☐
16	Ich wechsle/behalte meine Vermutung wegen des Experimentiergebnisses.	☐	☐	☐	☐	☐

Vielen Dank für Ihre Unterstützung!

B.8.2 *Fragebogen zur empirischen Testevaluation*

R

HUMBOLDT-UNIVERSITÄT ZU BERLIN

Lehrstuhl für Didaktik der Physik

Fragebogen zur Erfassung von Begründungen beim Experimentieren

Mit diesem Fragebogen möchten wir herausfinden, was du beim Experimentieren denkst. Es handelt sich hierbei **nicht** um eine Klassenarbeit oder einen Test. Es gibt **keine** falschen Antworten. Der Fragebogen ist vollständig **anonym**. Kreuze einfach die Antworten an, die für dich am besten passen.

Anleitung
- Lies bitte jede Frage sorgfältig durch und beantworte sie so genau wie möglich.
- Pro Frage ist nur ein Kreuz als Antwort erlaubt.
- Wenn du dich vertan hast, male bitte das Kästchen ganz schwarz aus: ■. Kreuze dann ein anderes an.
- Beantworte bitte alle Fragen.
- Bearbeite zunächst bitte nur die erste Seite.
- Führe dann das Experiment durch.

Los geht's!

Vielen Dank, dass du mitmachst!

Prof. Burkhard Priemer, Jakob Bar und Tobias Ludwig

Zunächst benötigen wir ein paar Daten zu deiner Person:

Wie alt bist du?

☐☐

Deine letzte Zeugnisnote in Physik war eine...

☐

Du bist ein...

Junge ☐
Mädchen ☐

HUMBOLDT-UNIVERSITÄT ZU BERLIN

Lehrstuhl für Didaktik der Physik

Einleitung:
Ein Fadenpendel besteht aus einer Masse, die an einem frei beweglichen Faden aufgehängt ist, ähnlich einer Schaukel auf dem Spielplatz. Ein Fadenpendel kann durch die folgenden Größen beschrieben werden:

T Schwingungsdauer: die Zeit, die ein Pendel benötigt, um eine komplette Schwingung durchzuführen, z.B. von der Ruhelage ganz nach links, dann ganz nach rechts und wieder zurück zur Ruhelage

l Länge des Fadens, an dem die Masse aufgehängt ist

m Die angehangene Masse

φ Auslenkung: der Winkel, um den das Pendel zu Beginn der Schwingung aus der Ruhelage gehoben wird

---

1. Es ist anzunehmen, dass Schwingungsdauer, Fadenlänge, Masse und Auslenkung in irgendeiner Weise miteinander zusammenhängen. Wir wollen im Folgenden nur untersuchen, wie die **Masse des Pendels** mit der **Schwingungsdauer** zusammenhängt. Stelle dazu zunächst deine eigene Vermutung auf:

Vergrößern der Pendelmasse bewirkt, dass die Schwingungsdauer…

…größer wird ☐
…gleich bleibt ☐
…kleiner wird ☐

---

2. Begründe **kurz** deine Vermutung!

Warte nun kurz, bis deine Mitschülerinnen und Mitschüler auch an dieser Stelle des Fragebogens sind.

AUFGABE:

Überprüfe nun deine gerade aufgestellte Vermutung mit dem Experiment!

Bitte bespreche dich dabei nicht mit deinen Nachbarn.
Blättere bitte erst um, wenn du mit dem Experiment fertig bist !

3. Mit welchem Experiment hast du gearbeitet?

Realexperiment ☐
Computerexperiment ☐

---

4. Behältst du deine Vermutung von Beginn bei?

Ja ☐ ⟶ Gehe zu Frage 5
Nein ☐

↳ Wenn du „Nein“ angekreuzt hast: Wie lautet deine neue Vermutung?
Vergrößern der Pendelmasse bewirkt, dass die Schwingungsdauer…

…größer wird ☐
…gleich bleibt ☐
…kleiner wird ☐

---

5. Wie sicher bist du dir, dass es richtig ist, deine Vermutung beizubehalten/zu verwerfen?

sehr unsicher	etwas unsicher	relativ sicher	sehr sicher
☐	☐	☐	☐

---

6. Du hast gerade deine Vermutung mit dem Experiment überprüft. In Frage 4 hast du dich entschieden, ob du deine Vermutung von Beginn beibehältst oder verwirfst. Du findest im Folgenden eine Reihe von Aussagen, die sich genau auf **diese Entscheidung** beziehen.

AUFGABE:

Denke einen Moment über jede der folgenden Aussage nach. Kreuze dann bitte an, wie sehr diese Aussagen **bei deiner Entscheidung eine Rolle gespielt haben**. Du kannst jede Aussage mit „trifft gar nicht zu“, „trifft wenig zu“, trifft teils/teils zu“, „trifft ziemlich zu“ und „trifft völlig zu“ bewerten.

Denke dabei immer an das Experiment, dass du **gerade eben** durchgeführt hast.

Ganz wichtig: Es gibt keine **richtigen** oder **falschen Antworten.**

#		trifft gar nicht zu	trifft wenig zu	trifft teils/teils zu	trifft ziemlich zu	trifft völlig zu
1	Ich denke erst über die Messdaten nach, bevor ich eine Entscheidung treffe.	☐	☐	☐	☐	☐
2	Bei meiner Entscheidung zum Beibehalten/Verwerfen meiner Vermutung berücksichtige ich, dass das Experiment an einer Universität entwickelt wurde.	☐	☐	☐	☐	☐
3	Ich ziehe die Schlussfolgerungen aus dem Experiment eher aufgrund meiner Gefühle.	☐	☐	☐	☐	☐
4	Die Messdaten sind so eindeutig, dass ich das Beibehalten/Verwerfen meiner Vermutung damit stützen kann.	☐	☐	☐	☐	☐
5	Bei der Entscheidung zum Beibehalten/Verwerfen meiner Vermutung habe ich deutlich auf meine Gefühle gehört.	☐	☐	☐	☐	☐
6	Bei diesem Experiment ist es wichtig, die Messung mehrfach zu wiederholen, um sichere Aussagen machen zu können.	☐	☐	☐	☐	☐
7	Anhand meiner Messwerte sehe ich, dass es richtig ist, meine Vermutung beizubehalten/zu verwerfen.	☐	☐	☐	☐	☐
8	Ich denke gar nicht nach, wenn ich mich für eine Vermutung entscheide.	☐	☐	☐	☐	☐
9	Beim Messen beobachtete ich Ungenauigkeiten. Das berücksichtige ich in hohem Maße bei meiner Entscheidung.	☐	☐	☐	☐	☐
10	Ich wechsle/beibehalte meine Vermutung wegen der Messwerte auf der Stoppuhr.	☐	☐	☐	☐	☐
11	Bei der Entscheidung berücksichtige ich stark mein Gefühl.	☐	☐	☐	☐	☐
12	Anhand der Messwerte überprüfe ich, ob ich meine Vermutung wechseln oder beibehalten soll.	☐	☐	☐	☐	☐

#		trifft gar nicht zu	trifft wenig zu	trifft teils/teils zu	trifft ziemlich zu	trifft völlig zu
13	Ich bin unsicher, ob ich meine Vermutung wechsle/beibehalte, weil die Messwerte leichte Abweichungen haben.	☐	☐	☐	☐	☐
14	Ich habe eher geraten als richtig über die experimentellen Beobachtungen nachgedacht.	☐	☐	☐	☐	☐
15	Bei meiner Entscheidung berücksichtige ich auch, dass das Experiment ungenau ist.	☐	☐	☐	☐	☐
16	So ein Experiment von einer Universität ist ja meistens richtig. Dies berücksichtige ich bei meiner Entscheidung.	☐	☐	☐	☐	☐
17	Die Ungenauigkeiten beim Experimentieren erschweren meine Entscheidung.	☐	☐	☐	☐	☐
18	Die Gründe für meine Entscheidung zum Beibehalten/Verwerfen meiner Vermutung sind mir nicht richtig bewusst.	☐	☐	☐	☐	☐
19	Ich berücksichtige bei meiner Entscheidung, dass das Messen mit der Stoppuhr nicht ganz genau ist.	☐	☐	☐	☐	☐
20	Ich habe mich spontan und schnell entschieden, ob ich meine Vermutung wechsle/beibehalte.	☐	☐	☐	☐	☐
21	Wegen dem, was ich beim Messen herausgefunden habe, wechsle/verwerfe ich meine Vermutung.	☐	☐	☐	☐	☐
22	Bei dieser Entscheidung zum Beibehalten/Verwerfen habe ich aus dem Bauch heraus entschieden.	☐	☐	☐	☐	☐
23	Ich wechsle/behalte meine Vermutung wegen des Experimentiergebnisses.	☐	☐	☐	☐	☐
24	Bei der Entscheidung berücksichtige ich, dass das Experiment ja schon einmal von anderen überprüft wurde.	☐	☐	☐	☐	☐
25	Ich habe mich gefühlsmäßig für die Antwort entschieden, die mir am meisten zusagt.	☐	☐	☐	☐	☐

#		trifft gar nicht zu	trifft wenig zu	trifft teils/teils zu	trifft ziemlich zu	trifft völlig zu
26	Bei diesem Experiment gibt es Messungenauigkeiten, die ich bei meiner Schlussfolgerung berücksichtige.	☐	☐	☐	☐	☐
27	Ich habe ein Gespür dafür, dass meine Vermutung richtig/falsch ist	☐	☐	☐	☐	☐
28	Ich wechsle/beibehalte meine Vermutung, weil ich das durch die gemessenen Zeiten begründen kann.	☐	☐	☐	☐	☐
29	Mein Gefühl sagt mir, dass ich meine Vermutung in diesem Fall beibehalten/wechseln sollte.	☐	☐	☐	☐	☐
30	Bei meiner Entscheidung habe ich gar nicht richtig nachgedacht, ich habe eher geraten.	☐	☐	☐	☐	☐
31	Ich habe die Messungenauigkeiten beim Experimentieren in hohem Maße berücksichtigt.	☐	☐	☐	☐	☐
32	Diese Ergebnisse haben mir gezeigt, dass es richtig ist, meine Vermutung beizubehalten/zuverwerfen.	☐	☐	☐	☐	☐
33	Bei meiner Entscheidung spielen Gefühle eine große Rolle.	☐	☐	☐	☐	☐
34	Das Experiment ist ja schon einmal geprüft worden, deshalb vertraue ich darauf, dass es fehlerfrei ist. Dies berücksichtige ich bei meiner Entscheidung.	☐	☐	☐	☐	☐
35	Bei meiner Entscheidung berücksichtige ich, dass das Experiment professionell aussah, weil es von Fachleuten aufgebaut wurde.	☐	☐	☐	☐	☐
36	Bevor ich die Entscheidung zum Beibehalten/Verwerfen meiner Vermutung treffe, denke ich zunächst einmal gründlich über die Messdaten nach.	☐	☐	☐	☐	☐
37	Bei meiner Entscheidung berücksichtige ich, dass das Experiment, das die Wissenschaftler hier mitgebracht haben, bestimmt nicht falsch sein wird.	☐	☐	☐	☐	☐
38	Bei der Entscheidung zum Beibehalten/Verwerfen meiner Vermutung verlasse ich mich auf mein Gespür.	☐	☐	☐	☐	☐

#		trifft gar nicht zu	trifft wenig zu	trifft teils/teils zu	trifft ziemlich zu	trifft völlig zu
39	Ich stütze meine Entscheidung zum Beibehalten/Wechseln meiner Vermutung auf die Fakten, die das Experiment mir liefert.	☐	☐	☐	☐	☐
40	Ich höre stark auf mein Bauchgefühl, wenn ich aus dem Experiment Schlüsse ziehe.	☐	☐	☐	☐	☐
41	Ohne großes Nachdenken erscheint es mir in diesem Fall richtig, meine Vermutung beizubehalten/zu verwerfen.	☐	☐	☐	☐	☐
42	Die Ergebnisse zeigen mir eindeutig, dass ich meine Vermutung verwerfen/beibehalten muss.	☐	☐	☐	☐	☐
43	Bei meiner Entscheidung zum Beibehalten/Verwerfen meiner Vermutung berücksichtige ich, dass die Messwerte deutlich schwanken.	☐	☐	☐	☐	☐
44	Meine Entscheidung, die Vermutung beizubehalten/zu verwerfen ist eine Reaktion auf die Messdaten.	☐	☐	☐	☐	☐
45	Bei der Entscheidung zum Wechseln/Beibehalten meiner Vermutung berücksichtige ich, dass man immer Ungenauigkeiten beim Experimentieren hat.	☐	☐	☐	☐	☐
46	Die Messdaten spielen bei meiner Entscheidung die größte Rolle.	☐	☐	☐	☐	☐
47	Ich berücksichtige, dass das Experiment von Wissenschaftlern mitgebracht wurde, die an der Humboldt-Universität arbeiten.	☐	☐	☐	☐	☐
48	Ich entscheide mich für/gegen meine Vermutung nur anhand meiner experimentellen Beobachtungen.	☐	☐	☐	☐	☐
49	Bei meiner Entscheidung berücksichtige ich, dass die Fachleute bestimmt keine Fehler im Aufbau des Experiments gemacht haben.	☐	☐	☐	☐	☐
50	Die Auswertung meiner Messung ist der Grund für meine Entscheidung zum Beibehalten/Wechseln meiner Vermutung.	☐	☐	☐	☐	☐
51	Richtig entschieden habe ich mich nicht. Ich habe nur geraten.	☐	☐	☐	☐	☐

#		trifft gar nicht zu	trifft wenig zu	trifft teils/teils zu	trifft ziemlich zu	trifft völlig zu
52	Die Messfehler waren so groß, dass sie bei meiner Entscheidung eine entscheidende Rolle gespielt haben.	☐	☐	☐	☐	☐
53	Bei meiner Entscheidung berücksichtige ich, dass dieses Experiment von den Fachleuten bestimmt richtig funktioniert.	☐	☐	☐	☐	☐
54	Ich berücksichtige bei meiner Entscheidung, dass die Menschen, die das Experiment aufgebaut haben, mehr Fachwissen haben als ich.	☐	☐	☐	☐	☐
55	Meine Messdaten sind der Grund, warum ich meine Vermutung verwerfe/beibehalte.	☐	☐	☐	☐	☐
56	Ich vertraue demjenigen, der das Experiment aufgebaut hat. Das berücksichtige ich bei meiner Entscheidung.	☐	☐	☐	☐	☐
57	Um für mich diese Entscheidung zum Beibehalten/Verwerfen meiner Vermutung zu begründen, denke ich besonders intensiv über die Messdaten nach.	☐	☐	☐	☐	☐
58	Bei der Entscheidung habe ich mich ganz auf mein Gefühl verlassen.	☐	☐	☐	☐	☐
59	Ich habe beim Stoppen Zeiten gemessen und diese Messwerte veranlassen mich, meine Vermutung beizubehalten/zu wechseln.	☐	☐	☐	☐	☐
60	Wenn ich die Messdaten interpretiere, muss ich auch berücksichtigen, dass das Experiment etwas ungenau ist.	☐	☐	☐	☐	☐
61	Die Messwerte sind ausreichend um meine Vermutung zu wechseln/beizubehalten.	☐	☐	☐	☐	☐
62	Bei meiner Entscheidung berücksichtige ich, dass das Experiment bereits von Wissenschaftlern erdacht und überprüft wurde.	☐	☐	☐	☐	☐
63	Ich berücksichtige bei meiner Entscheidung, dass das Messen mit der Stoppuhr nicht ganz genau ist.	☐	☐	☐	☐	☐

Vielen Dank für deine Unterstützung!

C

ANHANG METHODEN HAUPTSTUDIE

C.1 ÜBERSICHT ÜBER DIE ITEMTEXTE DER VERWENDETEN SKALEN

Im Folgenden sind die Itemtexte der verwendenten Skalen dargestellt. Methodologische Überlegungen und Entscheidungen zur Auswahl dieser Skalen sowie die entsprechenden Referenzen sind in Abschnitt 4.3.3 dargestellt.

Die Items sind hier sortiert aufgeführt. Im Fragebogen wurde entsprechend der Reihenfolge der Instrumente (siehe Abschnitt 4.5) eine zufällige Sortierung verwendet (vgl. Abschnitt C.2).

C.1.1 *Argumentationstest*

ARGUMENTKATEGORIE INTUITION

`i1.03` Ich höre stark auf mein Bauchgefühl, wenn ich aus dem Experiment Schlüsse ziehe.

`i1.09` Bei der Entscheidung berücksichtige ich stark mein Gefühl.

`i1.16` Bei meiner Entscheidung spielen Gefühle eine große Rolle.

`i1.18` Ich habe mich gefühlsmäßig für die Antwort entschieden, die mir am meisten zusagt.

`i1.21` Bei der Entscheidung habe ich mich ganz auf mein Gefühl verlassen.

ARGUMENTKATEGORIE EXPERTENWISSEN

`i6.05` Bei meiner Entscheidung berücksichtige ich, dass dieses Experiment von den Fachleuten bestimmt richtig funktioniert.

`i6.08` Bei meiner Entscheidung berücksichtige ich, dass das Experiment bereits von Wissenschaftlern erdacht und überprüft wurde.

`i6.11` Bei meiner Entscheidung berücksichtige ich, dass die Fachleute bestimmt keine Fehler im Aufbau des Experiments gemacht haben.

`i6.15` Ich berücksichtige, dass das Experiment von Wissenschaftlern mitgebracht wurde, die an der Humboldt-Universität arbeiten.

i6.16 Ich vertraue demjenigen, der das Experiment aufgebaut hat. Das berücksichtige ich bei meiner Entscheidung.

ARGUMENTKATEGORIE MESSUNSICHERHEITEN (EXPLIZIT)

i7b.04 Bei diesem Experiment gibt es Messungenauigkeiten, die ich bei meiner Schlussfolgerung berücksichtige.

i7b.08 Die Ungenauigkeiten beim Experimentieren erschweren meine Entscheidung.

i7b.16 Bei der Entscheidung zum Wechseln/Beibehalten meiner Vermutung berücksichtige ich, dass man immer Ungenauigkeiten beim Experimentieren hat.

i7b.17 Bei meiner Entscheidung berücksichtige ich auch, dass das Experiment ungenau ist.

i7b.22 Ich habe die Messungenauigkeiten beim Experimentieren in hohem Maße berücksichtigt.

ARGUMENTKATEGORIE EVIDENZ

i8.05 Meine Messdaten sind der Grund, warum ich meine Vermutung verwerfe/beibehalte.

i8.09 Meine Entscheidung, die Vermutung beizubehalten/zu verwerfen ist eine Reaktion auf die Messdaten.

i8.11 Um für mich diese Entscheidung zum Beibehalten/Verwerfen meiner Vermutung zu begründen, denke ich besonders intensiv über die Messdaten nach.

i8.14 Die Messdaten spielen bei meiner Entscheidung die größte Rolle.

i8.18 Die Auswertung meiner Messung ist der Grund für meine Entscheidung zum Beibehalten/Wechseln meiner Vermutung.

C.1.2 *Kognitionsbedürfnis*

Die Items nfc4, nfc6 , nfc7, nfc8, nfc9, nfc10, nfc11, nfc12, nfc15, nfc16 sind negativ formuliert und wurden vor den Analysen invertiert.

nfc1 Die Aufgabe, neue Lösungen für Probleme zu finden, macht mir wirklich Spaß.

nfc2 Ich würde lieber eine Aufgabe lösen, die Intelligenz erfordert, schwierig und bedeutend ist, als eine Aufgabe, die zwar irgendwie wichtig ist, aber nicht viel Nachdenken erfordert.

`nfc3` Ich setze mir eher solche Ziele, die nur mit erheblicher geistiger Anstrengung erreicht werden können.

`nfc4` Die Vorstellung, mich auf mein Denkvermögen zu verlassen, um es zu etwas zu bringen, spricht mich nicht an.

`nfc5` Ich finde es besonders befriedigend, eine bedeutende Aufgabe abzuschließen, die viel Denken und geistige Anstrengung erfordert hat.

`nfc6` Ich denke lieber über kleine, alltägliche Vorhaben nach, als über langfristige.

`nfc7` Ich würde lieber etwas tun, das wenig Denken erfordert, als etwas, das mit Sicherheit meine Denkfähigkeit herausfordert.

`nfc8` Ich finde wenig Befriedigung darin, angestrengt und stundenlang nachzudenken.

`nfc9` In erster Linie denke ich, weil ich muss.

`nfc10` Ich trage nicht gerne die Verantwortung für eine Situation, die sehr viel Denken erfordert.

`nfc11` Denken entspricht nicht dem, was ich unter Spaß verstehe.

`nfc12` Ich versuche, Situationen vorauszuahnen und zu vermeiden, in denen die Wahrscheinlichkeit groß ist, dass ich intensiv über etwas nachdenken muss.

`nfc13` Ich habe es gern, wenn mein Leben voller kniffliger Aufgaben ist, die ich lösen muss.

`nfc14` Ich würde komplizierte Probleme einfachen Problemen vorziehen.

`nfc15` Es genügt mir, einfach die Antwort eines Problems zu verstehen.

`nfc16` Es genügt, daß etwas funktioniert, mir ist es egal, wie oder warum.

C.1.3 *Situationales Interesse*

EMOTIONALE KOMPONENTE

`si.emo.1` Die Experimentieraufgabe ist für mich unterhaltsam.

`si.emo.2` Ich finde die Experimentieraufgabe spannend.

`si.emo.3` Die Experimentieraufgabe macht mir Spaß.

AUFMERKSAMKEITSBEZOGENE KOMPONENTE

si.auf.1 Die Experimentieraufgabe weckt meine Neugier.

si.auf.2 Die Experimentieraufgabe fesselt meine Aufmerksamkeit.

si.auf.3 Ich konzentriere mich auf die Experimentieraufgabe.

EPISTEMISCHE KOMPONENTE

si.epi.1 Ich bin auf Themen gestoßen, zu denen ich gerne mehr Information hätte.

si.epi.2 Über Teile der Aufgabe möchte ich gerne mehr erfahren.

si.epi.3 Für mich haben sich neue Fragen ergeben, auf die ich gerne eine Antwort hätte.

si.epi.4 Ich möchte mich über die Inhalte der Experimentieraufgabe mit anderen unterhalten.

WERTBEZOGENE KOMPONENTE

si.wert.1 Die Experimentieraufgabe ist mir wichtig.

si.wert.2 Die Beschäftigung mit der Experimentieraufgabe ist für mich nützlich.

si.wert.3 Die Experimentieraufgabe ist für mich bedeutsam.

C.1.4 *Werteinschätzung der Naturwissenschaften*

SUBSKALA PERSÖNLICHE WERTVORSTELLUNGEN DER NATURWISSENSCHAFTEN

wdn.pers.1 Ich finde, dass Physik mir hilft, die Dinge um mich herum zu verstehen.

wdn.pers.2 Ich werde Physik in vielfältiger Weise nutzen, wenn ich erwachsen bin.

wdn.pers.3 Manche naturwissenschaftliche Konzepte helfen mir zu verstehen, wie ich mit anderen Menschen in Beziehung stehe.

wdn.pers.4 Wenn ich die Schule verlasse, werde ich viele Gelegenheiten haben, Wissen über Physik anzuwenden.

wdn.pers.5 Physikalisches Wissen ist für meinen Alltag wichtig.

SUBSKALA HANDLUNGSBEZOGENE WERTVORSTELLUNGEN DER NATURWISSENSCHAFTEN

`wdn.handlung.1` Das Treffen einer guten Entscheidung gleicht einem wissenschaftlichen Prozess.

`wdn.handlung.2` Physik hilft mir vernünftige Entscheidungen zu treffen.

`wdn.handlung.3` Das Sammeln von Informationen ist ein wichtiger Teil des Treffens von Entscheidungen.

`wdn.handlung.4` Mithilfe physikalischer Methoden lassen sich Dinge besser durchdenken.

`wdn.handlung.5` Physik kann mir helfen, in vielen Bereichen meines Lebens bessere Entscheidungen zu treffen.

C.1.5 *Fachwissen Mechanik*

Die Items des Fachwissentests Mechanik sind im Fragebogen der Hauptuntersuchung (Abschnitt C.2) ab S. 5 dargestellt. Wie in Abschnitt 4.3.3 dargelegt, wurden die Items des Fachwissenstests von Zander (2016) übernommen und angepasst. Der Vollständigkeit halber sind hier die ursprünglichen Quellen der Items zusätzlich angegeben: Die Items `K12P01` und `K06P04` stammen aus TIMSS 1995, Pop. 2 (IEA, 1997), `K06P03` und `B03F01` aus TIMSS 1999, Pop. 2 (IEA, 2001), `D04F01` aus TIMSS 2003, Pop. 2 (IEA, 2007), `K06F02` aus TIMSS 2007, Pop. 1 (IEA, 2009), `K09P01`, `K12L01` und `K06P02` aus Kirschner, Borowski und Fischer (2011) und `E04P01` aus Duit, von Zelewski und Heyner (1978). Die Items `K10P01` und `K06L01` sind motiviert von http://leifiphysik.de. Die Items `K02U01`, `K07P01`, `K13L01`, `K01P01`, `B02M01`, `B02F03`, `D02P01`, `D01P02`, `E06M01` und `E05L01` sind Eigenentwicklungen von Zander (2016).

Zur besseren Übersicht sind die Itemcodes des Mechaniktests hier in Stichworten beschrieben.

`K12P01` Wippe

`K02U01` Bleistift

`K06F02` Gewichtskraft

`K09P01` Parallelogramm

`K10P01` Lampe

`K12L01` Hebel

`K06P04` Apfel

K06L01 Astronaut

K07P01 Körper/Kräfte

K13L01 Flaschenzug

K01P01 Planet / Masse

K06P03 Rakete

K06P02 Federwaage Mond

B03F01 Weg-Zeit-Diegramm

B02M01 gleichförmige Bewegung

B02F03 Sportler

D04F01 Ballon

D02P01 Körper / Massen

D01P02 Druck

E06M01 Clowns

E04P01 Leistung

E05L01 Stabhochsprung

C.2 FRAGEBOGEN HAUPTUNTERSUCHUNG

HUMBOLDT-UNIVERSITÄT ZU BERLIN

Lehrstuhl für Didaktik der Physik

Studie zum Argumentieren beim Experimentieren

HAUPTTEST

Liebe Schülerin, lieber Schüler,

in dieser Studie möchten wir untersuchen, was du beim Experimentieren in Physik denkst.

Der Ablauf heute ist wie folgt: Zunächst wirst du einige Fragen zur Physik bearbeiten. Dann wirst du selbstständig experimentieren und im Anschluss wieder einige Fragen beantworten, die sich auf dein Experiment beziehen. Die gesamte Untersuchung dauert heute ca. 90 Minuten.

Nach den Sommerferien wird dir dein Physiklehrer bzw. deine Physiklehrerin noch einen kurzen Nachfolgetest mitbringen, der aber nur 10 Minuten dauern wird.

Die Teilnahme ist natürlich freiwillig. Du kannst die Teilnahme auch zu jeder Zeit abbrechen. Dieses Testheft ist außerdem pseudonymisiert, so dass wir nicht heraus finden können, welcher Teilnehmer welches Heft bearbeitet hat. Wir werden auch keine Daten an eure Lehrerin oder euren Lehrer weitergeben.

Wir würden uns freuen, wenn du an dieser Studie teilnehmen würdest. Mit deiner Teilnahme trägst du dazu bei, besser zu verstehen, *wie* Schülerinnen und Schüler beim Experimentieren Physik lernen!

Ganz wichtig: Dies ist kein Test oder eine Klassenarbeit! Es kann sein, dass du einige Fragen nicht sicher beantworten kannst – das ist völlig in Ordnung! ☺

Anleitung

- Lies bitte jede Frage sorgfältig durch und beantworte sie so genau wie möglich.
- Pro Frage ist nur ein Kreuz als Antwort erlaubt.
- Wenn du dich vertan hast, male bitte das Kästchen ganz schwarz aus: ■ Kreuze dann ein anderes an.
- Wenn du ein Kästchen ausgemalt hast und es danach doch ankreuzen möchtest, setze ein Kreuz unmittelbar daneben: x■
- Beantworte bitte alle Fragen.
- Wenn du an einer Stelle nicht weiter weißt, frage bitte die Versuchsleiter.

Vielen Dank, dass du mitmachst!

Tobias Ludwig Burkhard Priemer

Tobias Ludwig und Prof. Burkhard Priemer

Angaben zu deiner Person

1. Wie alt bist du? ☐☐

2. Welche Klassenstufe besuchst du? ☐

3. Was war deine letzte Zeugnisnote in Physik? ☐ *(Wenn du dich nicht mehr erinnern kannst, trage bitte nichts ein. Wenn du die Note auf dem kommenden Zeugnis bereits kennst trage bitte diese Note ein.)*

4. Du bist ein...

   Junge ☐
   Mädchen ☐

5. Damit wir diesen Fragebogen und den zweiten Fragebogen (nach den Sommerferien) miteinander verbinden können, benötigen wir ein Pseudonym:

   Wie lautet der zweite Buchstabe deines Vornamens? ☐
   (Beispiel: Du heißt „Peter", dann lautet der zweite Buchstabe „E")

   Wie lautet der erste Buchstabe des Vornamens deiner Mutter? ☐
   (Beispiel: Deine Mutter heißt „Anna", dann lautet der erste Buchstabe „A")

   Wie lauten die ersten zwei Ziffern des Geburtstages deiner Mutter? ☐☐
   (Beispiel: Deine Mutter hat am 02. Mai Geburtstag, dann lauten die ersten beiden Ziffern „02")

   Wie lauten die ersten zwei Buchstaben der Straße in der du wohnst? ☐☐
   (Beispiel: Du wohnst in der „Frankfurter Allee", dann lauten die ersten zwei Buchstaben „FR")

Welche Bedeutung hat Physik für dich? Wie gerne denkst du über anspruchsvolle Probleme nach?

In diesem Teil des Fragebogens wollen wir herausfinden, welche Bedeutung die Physik und Naturwissenschaften für dich haben und wie gerne du dich mit Denkaufgaben beschäftigst.

AUFGABE:

6. Denke einen Moment über jede der folgenden Aussage nach. Kreuze dann an, wie sehr du mit diesen Aussagen übereinstimmst. Du kannst jede Aussage mit „stimme gar nicht zu“, „stimme wenig zu“, „stimme teils/teils zu“, „stimme ziemlich zu“ und „stimme völlig zu“ bewerten. Ganz wichtig: Es gibt keine **richtigen** oder **falschen Antworten.**

   Wie sehr stimmst du mit den folgenden Aussagen überein?

#	stimme gar nicht zu	stimme wenig zu	stimme teils/teils zu	stimme ziemlich zu	stimme völlig zu
Das Treffen einer guten Entscheidung gleicht einem wissenschaftlichen Prozess.	☐	☐	☐	☐	☐
Manche naturwissenschaftliche Konzepte helfen mir zu verstehen, wie ich mit anderen Menschen in Beziehung stehe.	☐	☐	☐	☐	☐
Physik hilft mir vernünftige Entscheidungen zu treffen.	☐	☐	☐	☐	☐
Mithilfe physikalischer Methoden lassen sich Dinge besser durchdenken.	☐	☐	☐	☐	☐
Physikalisches Wissen ist für meinen Alltag wichtig.	☐	☐	☐	☐	☐
Ich werde Physik in vielfältiger Weise nutzen, wenn ich erwachsen bin.	☐	☐	☐	☐	☐
Physik kann mir helfen, in vielen Bereichen meines Lebens bessere Entscheidungen zu treffen.	☐	☐	☐	☐	☐
Wenn ich die Schule verlasse, werde ich viele Gelegenheiten haben, Wissen über Physik anzuwenden.	☐	☐	☐	☐	☐
Das Sammeln von Informationen ist ein wichtiger Teil des Treffens von Entscheidungen.	☐	☐	☐	☐	☐
Ich finde, dass Physik mir hilft, die Dinge um mich herum zu verstehen.	☐	☐	☐	☐	☐

#	stimme gar nicht zu	stimme wenig zu	stimme teils/teils zu	stimme ziemlich zu	stimme völlig zu
Die Aufgabe, neue Lösungen für Probleme zu finden, macht mir wirklich Spaß.	☐	☐	☐	☐	☐
Ich würde lieber eine Aufgabe lösen, die Intelligenz erfordert, schwierig und bedeutend ist, als eine Aufgabe, die zwar irgendwie wichtig ist, aber nicht viel Nachdenken erfordert.	☐	☐	☐	☐	☐
Ich setze mir eher solche Ziele, die nur mit erheblicher geistiger Anstrengung erreicht werden können.	☐	☐	☐	☐	☐
Die Vorstellung, mich auf mein Denkvermögen zu verlassen, um es zu etwas zu bringen, spricht mich nicht an.	☐	☐	☐	☐	☐
Ich finde es besonders befriedigend, eine bedeutende Aufgabe abzuschließen, die viel Denken und geistige Anstrengung erfordert hat.	☐	☐	☐	☐	☐
Ich denke lieber über kleine, alltägliche Vorhaben nach, als über langfristige.	☐	☐	☐	☐	☐
Ich würde lieber etwas tun, das wenig Denken erfordert, als etwas, das mit Sicherheit meine Denkfähigkeit herausfordert.	☐	☐	☐	☐	☐
Ich finde wenig Befriedigung darin, angestrengt und stundenlang nachzudenken.	☐	☐	☐	☐	☐
In erster Linie denke ich, weil ich muss.	☐	☐	☐	☐	☐
Ich trage nicht gerne die Verantwortung für eine Situation, die sehr viel Denken erfordert.	☐	☐	☐	☐	☐
Denken entspricht nicht dem, was ich unter Spaß verstehe.	☐	☐	☐	☐	☐
Ich versuche, Situationen vorauszuahnen und zu vermeiden, in denen die Wahrscheinlichkeit groß ist, dass ich intensiv über etwas nachdenken muss.	☐	☐	☐	☐	☐
Ich habe es gern, wenn mein Leben voller kniffliger Aufgaben ist, die ich lösen muss.	☐	☐	☐	☐	☐
Ich würde komplizierte Probleme einfachen Problemen vorziehen.	☐	☐	☐	☐	☐
Es genügt mir, einfach die Antwort eines Problems zu verstehen.	☐	☐	☐	☐	☐
Es genügt, dass etwas funktioniert, mir ist es egal, wie oder warum.	☐	☐	☐	☐	☐

Wie viel weißt du bereits über Physik?

In diesem Teil des Fragebogens möchten wir herausfinden, wie viel du über Mechanik bereits weißt. Aber wichtig: Das ist kein Test oder eine Klassenarbeit! Es ist hier ganz normal, dass du vielleicht ein paar Fragen nicht sicher beantworten kannst! Das ist überhaupt nicht schlimm!

AUFGABE:

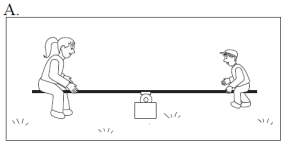
7. Bitte bearbeite die folgenden Aufgaben zur Physik!

K12P01

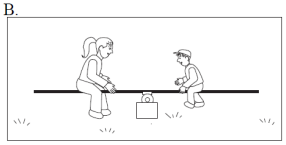
Ein Mädchen spielt mit ihrem kleinen Bruder auf einer Wippe. Das Mädchen wiegt 50 kg (Kilogramm), der Junge wiegt 25 kg.

Welches Bild zeigt die beste Position für das Mädchen, um mit seinem Bruder im Gleichgewicht zu sein?

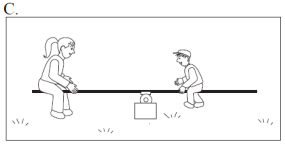
A.



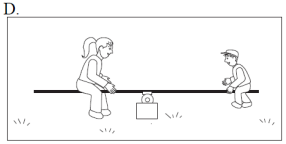
B.



C.



D.



☐ A

☐ B

☐ C

☐ D

K02U01

Was wird passieren, wenn du versuchst, mit deinen Händen einen Bleistift zu verbiegen?

☐ Der Bleistift wird seine Form keinesfalls verändern.

☐ Der Bleistift verbiegt sich zunächst und bricht, falls du zu stark drückst.

☐ Der Bleistift bricht, egal wie stark du drückst.

☐ Der Bleistift geht in jedem Fall in seine Ursprungsform zurück.

K06F02

In welchem Beispiel bewegt sich ein Objekt wegen der Gewichtskraft?

☐ Ein Mädchen schlägt einen Ball mit einem Schläger.

☐ Ein Junge schiebt eine Kiste über den Boden.

☐ Ein Mädchen hämmert einen Nagel in die Wand.

☐ Ein Junge fällt von einem Baum auf den Boden.

K09P01

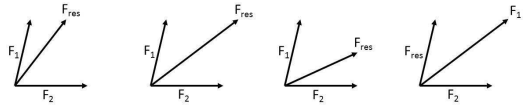
In welchem der folgenden Bilder wurde der resultierende Kraftpfeil F_{res} korrekt eingezeichnet?

☐ A

☐ B

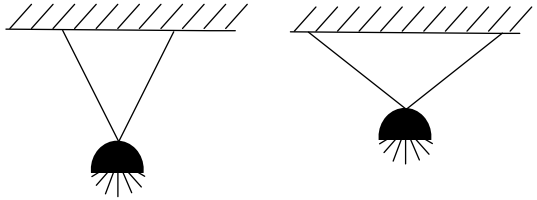
☐ C

☐ D



K10P01

Eine Lampe hängt an zwei Seilen. Auf die Lampe wirken die Gewichtskraft senkrecht nach unten, sowie die Seilkräfte in Richtung der Seile.



Wie ändern sich die Seilkräfte, wenn die Lampe höher gehängt wird?

☐ Je höher die Lampe gezogen wird, desto größer werden die Seilkräfte.

☐ An den Seilkräften ändert sich nichts.

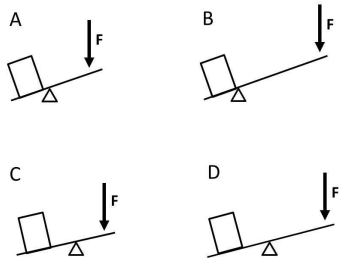
☐ Je höher die Lampe gezogen wird, desto kleiner werden die Seilkräfte.

☐ Entlang der Seile wirken dann keine Kräfte mehr.

K12L01

Eine sehr schwere Kiste soll mit Hilfe eines Hebels hochgehoben werden.

Welche der folgenden Hebelvorrichtungen (siehe Abbildung unten) muss man benutzen, wenn die Kraft F, die man ausüben muss, um den Ziegelstein anzuheben, möglichst gering sein soll?



☐ A

☐ B

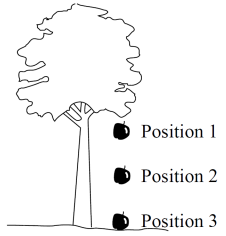
☐ C

☐ D

K06P04

Das Bild zeigt einen zu Boden fallenden Apfel.

In welcher oder in welchen der drei Positionen wirkt die Gewichtskraft auf den Apfel?



☐ Nur in Position 2.

☐ Nur in Position 1 und 2.

☐ Nur in Position 1 und 3.

☐ In Position 1, 2 und 3.

K06L01

Ein Astronaut sammelt auf dem Mond Gestein. Dieses Gestein erfährt eine Gewichtskraft von 16 N. Auf dem Mond beträgt der Ortsfaktor $g = 1{,}6\frac{N}{kg}$. (Hinweis: Vielleicht kennst du den Begriff Ortsfaktor auch unter dem Begriff „Schwerebeschleunigung“ oder „Fallbeschleunigung“).

Wie groß ist die Gewichtskraft des Gesteins auf der Erde $(g = 10\frac{N}{kg})$?

☐ 10 N

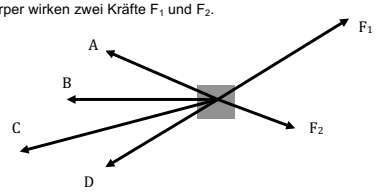
☐ 16 N

☐ 100 N

☐ 160 N

K07P01

Auf einen Körper wirken zwei Kräfte F_1 und F_2.



Welche der eingezeichneten Kräfte A, B, C oder D muss wirken, damit die auf den Körper wirkende resultierende Kraft den Betrag Null hat.

☐ Kraft A

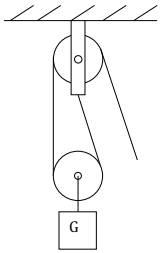
☐ Kraft B

☐ Kraft C

☐ Kraft D

K13L01

In der Abbildung ist ein Flaschenzug dargestellt.



Die angehangene Last G erfährt eine Gewichtskraft von 10 N. Zieht man das Seil 1 m, so…

☐ …muss man eine Kraft von 5 N aufwenden und die Last wird 0,5 m in die Höhe gezogen.

☐ …muss man eine Kraft von 5 N aufwenden und die Last wird 1 m in die Höhe gezogen.

☐ …muss man eine Kraft von 10 N aufwenden und die Last wird 0,5 m in die Höhe gezogen.

☐ …muss man eine Kraft von 10 N aufwenden und die Last wird 1 m in die Höhe gezogen.

K01P01

Ein Körper besitzt auf der Erde $(g = 10\frac{N}{kg})$ eine Masse von 50 kg.

Wie groß ist die Masse des Körpers auf einem Planeten mit $g = 20\frac{N}{kg}$?

☐ 25 kg

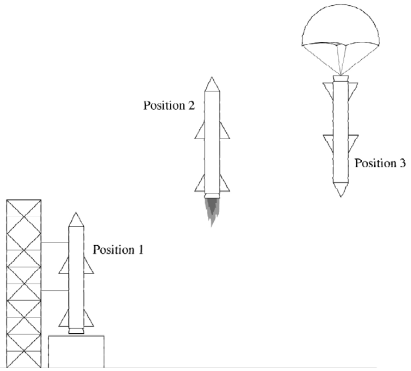
☐ 75 kg

☐ 50 kg

☐ 100 kg

K06P03

Die Abbildung zeigt eine Rakete, welche von der Erde abgeschossen wird und zurückkehrt.



In welcher der drei Positionen wirkt die Gewichtskraft auf die Rakete?

- ☐ nur in Position 3
- ☐ nur in Position 1 und 2
- ☐ nur in Position 2 und 3
- ☐ in Position 1, 2 und 3

K06P02

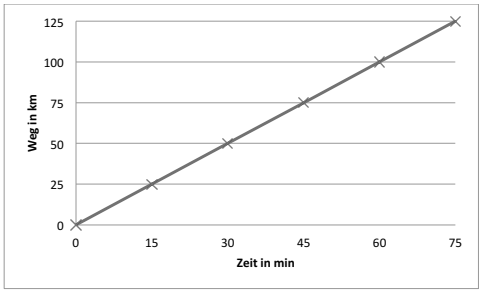
Du bekommst sieben Messingstücke mit Haken und eine Federwaage. Es wird behauptet, dass eines der Messingstücke so schwer ist wie die sechs anderen zusammen.

Könnte man die experimentelle Überprüfung der Behauptung auch auf dem Mond vornehmen?

- ☐ Ja, denn das Verhältnis der Gewichtskräfte bleibt gleich.
- ☐ Ja, denn die Gewichtskräfte bleiben gleich.
- ☐ Nein, denn es wird ein anderes Gewicht gemessen.
- ☐ Nein, die Federwaage funktioniert nur auf der Erde.

B03F01

Ein Auto fährt eine Straße entlang. Die Abbildung zeigt das zu dieser Bewegung gehörende Weg-Zeit-Diagramm.



Wie groß ist die Geschwindigkeit des Autos?

- ☐ 25 Kilometer pro Stunde
- ☐ 50 Kilometer pro Stunde
- ☐ 75 Kilometer pro Stunde
- ☐ 100 Kilometer pro Stunde

B02M01

Ein Körper bewegt sich geradlinig gleichförmig entlang eines Weges von 4 Metern in 8 Sekunden.

Wie weit bewegt sich der Körper, wenn er sich mit der dreifachen Geschwindigkeit 5 Sekunden lang bewegt?

- ☐ 2,5 m
- ☐ 4 m
- ☐ 7,5 m
- ☐ 30 m

B02F03

Ein Sportler lief 3000 m in genau 10 Minuten.

Was war seine durchschnittliche Geschwindigkeit in Metern pro Sekunde?

- ☐ 3
- ☐ 5
- ☐ 50
- ☐ 300

D04F01

Ein mit Heliumgas gefüllter Ballon wird los gelassen.

Welche der folgenden Aussagen erklärt am besten, warum der Heliumballon aufsteigt?

- ☐ Die Dichte von Helium ist kleiner als die Dichte der Luft.
- ☐ Der Luftwiderstand hebt den Ballon hoch.
- ☐ Die Schwerkraft wirkt nicht auf Heliumballons.
- ☐ Der Wind bläst den Ballon hoch.

D02P01

Die abgebildeten Körper besitzen das gleiche Volumen, aber unterschiedliche Massen.

Körper A (m = 0,8 kg) Körper B (m = 1,0 kg) Körper C (m = 1,2 kg)

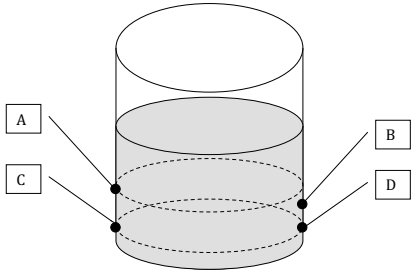
Was kann man über die Dichte der Körper sagen?

- ☐ Körper A besitzt die größte Dichte.
- ☐ Körper B besitzt die größte Dichte.
- ☐ Körper C besitzt die größte Dichte.
- ☐ Alle Körper besitzen die gleiche Dichte.

D01P02

In einem Glasgefäß befindet sich Wasser.

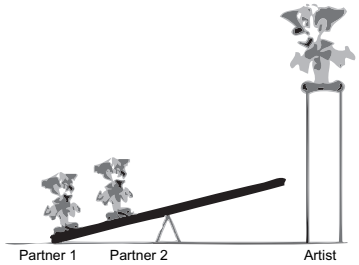
Was kann man über den Druck sagen, den das Wasser auf die markierten Stellen A, B, C, D der Gefäßwand ausübt?



- ☐ Der Druck ist an der Stelle A am größten.
- ☐ Der Druck an der Stelle A ist genauso groß wie der Druck an der Stelle B.
- ☐ Der Druck an der Stelle C ist genauso groß wie der Druck an der Stelle D.
- ☐ Der Druck ist überall gleich groß.

E06M01

Ein Artist mit der Masse 100 kg springt aus einer Höhe von 2 m auf ein Schleuderbrett (siehe Zeichnung).



Wie hoch werden seine zwei Partner geschleudert, wenn sie jeweils eine Masse von 50 kg haben?

- ☐ Partner 1 wird höher als Partner 2 geschleudert.
- ☐ Partner 2 wird höher als Partner 1 geschleudert.
- ☐ Beide werden gleich hoch geschleudert.
- ☐ Sie werden gar nicht hoch geschleudert.

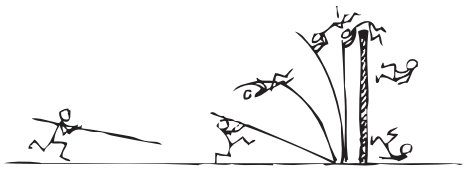
E04P01

Was versteht man in der Physik unter dem Begriff Leistung?

- ☐ Die Leistung gibt an, wie gut bei einem Vorgang die Energie in eine andere Energieform umgewandelt wird.
- ☐ Die Leistung gibt an, wie viel Energie bei einem Vorgang umgewandelt wird.
- ☐ Die Leistung gibt an, wie viel Energie bei einem Vorgang in einer bestimmten Zeit umgewandelt wird.
- ☐ Die Leistung gibt an, wie viel Energie bei einem Vorgang mit der nötigen Zeit multipliziert wird.

E05L01

Im Bild ist der Bewegungsablauf beim Stabhochsprung dargestellt.



Durch den Anlauf erhält der 70 kg wiegende Läufer eine Bewegungsenergie von 3500 J.

Welche Höhe kann er in etwa gewinnen, wenn er seine gesamte Bewegungsenergie in Höhenenergie umwandelt?

- ☐ 3 m
- ☐ 5 m
- ☐ 7 m
- ☐ 9 m

Eine Aufgabe zum Experimentieren

Marie ist 15 Jahre alt und geht in die 9. Klasse. Nach dem Frühstück liest sie manchmal die Nachrichten. Wirklich erstaunt ist sie über die folgende Meldung (das Foto rechts ist eine Vergrößerung der Abbildung aus dem Zeitungsausschnitt):

Mindener Tageblatt

Unabhängige, überparteiliche Zeitung

START | LOKALES | LOKALSPORT | WELTNEWS | MAGAZIN | MEINUNG | SERVICE

MINDEN | PORTA WESTFALICA | PETERSHAGEN | HILLE | UCHTE | KULTUR | WIRTSCHAFT | RE

Startseite » Lokales » Minden » Kletterer springen von Mindener Glacisbrücke...

Artikel vorlesen 01.02.2014

Kletterer springen von Mindener Glacisbrücke

Mindener veröffentlichen Video von Aktion auf Youtube / Stadt: Illegaler Flashmob

VON MAX STARK UND DANIEL MEHLKOPF

Minden (dm/ms). Mit einem lauten Schrei stürzt sich der Mann von der Glacisbrücke. Nur ein Seil schützt ihn vor dem Aufprall. Nach einem kurzen Fall fängt die Sicherung ihn auf, er schwingt knapp über der Wasseroberfläche wie ein Pendel hin und her.

Ausgelassen schwingen die Kletterer unter der Brücke hin und her. | Foto: Paul Olfermann

Ein kürzlich im Online-Portal Youtube veröffentlichtes Video dokumentiert diese Aktion von vier Klettersportlern im letzten Sommer. Mit einem Kletterseil lassen sich die Sportbegeisterten von der Glacisbrücke fallen. Ihre waghalsigen Sprünge filmen sie aus mehreren Perspektiven. Besonders ins Auge fallen die mit einer Action Cam aufgenommenen Sprungszenen. Dabei halten die Springer eine an einem Teleskopstab befestigte Kamera fest und springen mit dieser dann von der Brücke.

„Wow, das macht bestimmt Spaß!" Auf dem Bild sieht sie einen Springer am Seil durch die Luft hin und her pendeln. „Was passiert eigentlich, wenn man sich alleine nicht traut und zu zweit pendeln würde? Dauert es dann vielleicht länger oder kürzer, bis man einmal hin und wieder zurück geschwungen ist?" fragt sich Marie.

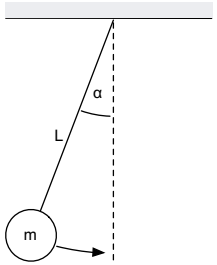
Man kann den Springer am Seil als ein Fadenpendel betrachten. Ein Fadenpendel besteht aus einer Masse, die an einem frei beweglichen Faden aufgehängt ist. Diese angehangene Masse entspricht der Masse Person, der Faden entspricht dem Seil. Ein Fadenpendel kann durch die folgenden Größen beschrieben werden:

T Schwingungsdauer: die Zeit, die ein Pendel benötigt, um eine komplette Schwingung durchzuführen, z. B. von der Ruhelage (Seil hängt senkrecht nach unten) zunächst ganz nach rechts, dann ganz nach links und wieder zurück zur Ruhelage. Je größer also die Schwingungsdauer, desto länger benötigt das Pendel für eine komplette Schwingung.

L Länge des Fadens, an dem die Masse aufgehängt ist

m Die angehangene Masse

α Auslenkung: der Winkel, um den das Pendel zu Beginn der Schwingung aus der Ruhelage gehoben wird



Es ist anzunehmen, dass Schwingungsdauer, Fadenlänge, Masse und Auslenkung in irgendeiner Weise miteinander zusammenhängen.

8. Marie weiß, dass sich die Masse des Pendels ändert, wenn z. B. zwei Personen anstatt einer Person zusammen schwingen. Marie möchte nun ein Experiment machen, um herauszufinden, wie sich die Schwingungsdauer ändert, wenn man die Masse ändert. Vor dem Experiment stellt sie zunächst eine Vermutung auf. Dabei berücksichtigt sie Effekte der Luftreibung und ganz große Winkel (größer als 90°) nicht.

   Stelle bitte auch du eine Vermutung auf, wie die Masse des Pendels mit der Schwingungsdauer zusammenhängt:

   Vergrößern der Pendelmasse bewirkt, dass die Schwingungsdauer, also die Zeit für eine komplette Schwingung, …

   …größer wird. ☐
   …gleich bleibt. ☐
   …kleiner wird. ☐

9. Begründe kurz deine Vermutung!

_____

_____

_____

_____

AUFGABE:

10. Denke einen Moment über jede der folgenden Aussage nach. Kreuze dann an, wie sehr diese Aussagen auf dich zutreffen. Du kannst jede Aussage mit „trifft gar nicht zu", „trifft wenig zu", „trifft teils/teils zu", „trifft ziemlich zu" und „trifft völlig zu" bewerten.

    Wie sehr treffen die folgenden Aussagen jetzt gerade auf dich zu?

#	trifft gar nicht zu	trifft wenig zu	trifft teils/teils zu	trifft ziemlich zu	trifft völlig zu
Ich möchte mich über die Inhalte der Experimentieraufgabe mit anderen unterhalten.	☐	☐	☐	☐	☐
Ich bin bei der Experimentieraufgabe auf Themen gestoßen, zu denen ich gerne mehr Information hätte.	☐	☐	☐	☐	☐

#	trifft gar nicht zu	trifft wenig zu	trifft teils/teils zu	trifft ziemlich zu	trifft völlig zu
Die Experimentieraufgabe ist für mich unterhaltsam.	☐	☐	☐	☐	☐
Über Teile der Experimentiertaufgabe möchte ich gerne mehr erfahren.	☐	☐	☐	☐	☐
Für mich haben sich durch die Experimentieraufgabe neue Fragen ergeben, auf die ich gerne eine Antwort hätte.	☐	☐	☐	☐	☐
Die Experimentieraufgabe weckt meine Neugier.	☐	☐	☐	☐	☐
Die Experimentieraufgabe macht mir Spaß.	☐	☐	☐	☐	☐
Die Experimentieraufgabe ist für mich bedeutsam.	☐	☐	☐	☐	☐
Ich konzentriere mich auf die Experimentieraufgabe.	☐	☐	☐	☐	☐
Die Experimentieraufgabe ist mir wichtig.	☐	☐	☐	☐	☐
Ich finde die Experimentieraufgabe spannend.	☐	☐	☐	☐	☐
Die Experimentieraufgabe fesselt meine Aufmerksamkeit.	☐	☐	☐	☐	☐
Die Beschäftigung mit der Experimentieraufgabe ist für mich nützlich.	☐	☐	☐	☐	☐

Warte nun kurz, bis deine Mitschülerinnen und Mitschüler ebenfalls an dieser Stelle des Fragebogens angekommen sind. Die Versuchsleiter geben dir Bescheid, wenn du weiter machen kannst.

AUFGABE:

11. Überprüfe nun deine gerade aufgestellte Vermutung mit dem Experiment!
    Bitte bespreche dich dabei nicht mit deinen Nachbarn.

Hier ist Platz für Notizen:

12. Mit welchem Experiment hast du gearbeitet?

Realexperiment ☐
Computerexperiment ☐

---

13. Behältst du deine Vermutung von Beginn bei?

Ja ☐ ⟶ Gehe zu Frage 14
Nein ☐

↳ Wenn du „Nein“ angekreuzt hast: Wie lautet deine neue Vermutung?
Vergrößern der Pendelmasse bewirkt, dass die Schwingungsdauer…

…größer wird ☐
…gleich bleibt ☐
…kleiner wird ☐

---

14. Wie sicher bist du dir, dass es richtig ist, deine Vermutung beizubehalten/zu verwerfen?

sehr unsicher	etwas unsicher	relativ sicher	sehr sicher
☐	☐	☐	☐

AUFGABE:

15. Du hast gerade deine Vermutung mit dem Experiment überprüft. Auf der Seite zuvor hast du dich entschieden, ob du deine Vermutung von Beginn beibehältst oder verwirfst. Du findest im Folgenden eine Reihe von Aussagen, die sich genau auf **diese Entscheidung** beziehen. Denke einen Moment über jede der folgenden Aussage nach. Du kannst jede Aussage mit „trifft gar nicht zu“, „trifft wenig zu“, trifft teils/teils zu“, „trifft ziemlich zu“ und „trifft völlig zu“ bewerten.

Kreuze bitte an wie sehr diese Aussagen bei deiner Entscheidung gerade eine Rolle gespielt haben. Denke dabei immer an das Experiment, dass du **gerade eben** durchgeführt hast. Ganz wichtig: Es gibt keine **richtigen** oder **falschen Antworten.**

#	trifft gar nicht zu	trifft wenig zu	trifft teils/teils zu	trifft ziemlich zu	trifft völlig zu
Bei der Entscheidung habe ich mich ganz auf mein Gefühl verlassen.	☐	☐	☐	☐	☐
Bei meiner Entscheidung berücksichtige ich auch, dass das Experiment ungenau ist.	☐	☐	☐	☐	☐
Die Auswertung meiner Messung ist der Grund für meine Entscheidung zum Beibehalten/Wechseln meiner Vermutung.	☐	☐	☐	☐	☐
Ich vertraue demjenigen, der das Experiment aufgebaut hat. Das berücksichtige ich bei meiner Entscheidung.	☐	☐	☐	☐	☐
Ich habe mich gefühlsmäßig für die Antwort entschieden, die mir am meisten zusagt.	☐	☐	☐	☐	☐
Die Messdaten spielen bei meiner Entscheidung die größte Rolle.	☐	☐	☐	☐	☐
Bei der Entscheidung berücksichtige ich stark mein Gefühl.	☐	☐	☐	☐	☐
Bei meiner Entscheidung berücksichtige ich, dass dieses Experiment von den Fachleuten bestimmt richtig funktioniert.	☐	☐	☐	☐	☐
Ich habe die Messungenauigkeiten beim Experimentieren in hohem Maße berücksichtigt.	☐	☐	☐	☐	☐
Meine Entscheidung, die Vermutung beizubehalten/zu verwerfen ist eine Reaktion auf die Messdaten.	☐	☐	☐	☐	☐

#		trifft gar nicht zu	trifft wenig zu	trifft teils/teils zu	trifft ziemlich zu	trifft völlig zu
	Bei meiner Entscheidung berücksichtige ich, dass das Experiment bereits von Wissenschaftlern erdacht und überprüft wurde.	☐	☐	☐	☐	☐
	Bei diesem Experiment gibt es Messungenauigkeiten, die ich bei meiner Schlussfolgerung berücksichtige.	☐	☐	☐	☐	☐
	Meine Messdaten sind der Grund, warum ich meine Vermutung verwerfe/beibehalte.	☐	☐	☐	☐	☐
	Bei meiner Entscheidung spielen Gefühle eine große Rolle.	☐	☐	☐	☐	☐
	Bei der Entscheidung zum Wechseln/Beibehalten meiner Vermutung berücksichtige ich, dass man immer Ungenauigkeiten beim Experimentieren hat.	☐	☐	☐	☐	☐
	Bei meiner Entscheidung berücksichtige ich, dass die Fachleute bestimmt keine Fehler im Aufbau des Experiments gemacht haben.	☐	☐	☐	☐	☐
	Ich höre stark auf mein Bauchgefühl, wenn ich aus dem Experiment Schlüsse ziehe.	☐	☐	☐	☐	☐
	Um für mich diese Entscheidung zum Beibehalten/Verwerfen meiner Vermutung zu begründen, denke ich besonders intensiv über die Messdaten nach.	☐	☐	☐	☐	☐
	Die Ungenauigkeiten beim Experimentieren erschweren meine Entscheidung.	☐	☐	☐	☐	☐
	Ich berücksichtige, dass das Experiment von Wissenschaftlern mitgebracht wurde, die an der Humboldt-Universität arbeiten.	☐	☐	☐	☐	☐

Vielen Dank für deine Unterstützung!

C.3 FRAGEBOGEN FOLLOW-UP-ERHEBUNG

HUMBOLDT-UNIVERSITÄT ZU BERLIN

Lehrstuhl für Didaktik der Physik

Studie zum Argumentieren beim Experimentieren

2. Teil

Liebe Schülerin, lieber Schüler,

vor den Sommerferien hast du an einer Studie teilgenommen, in der wir untersucht haben, wie du mit Daten und Beobachtungen umgehst, die du beim Experimentieren gesammelt hast. Du hast dazu auch ein Experiment zum Fadenpendel durchgeführt.

Wie angekündigt möchten wir dich heute noch einmal kurz befragen. Dieser Fragebogen dauert lediglich etwa 5 min.

Anleitung

- Lies bitte jede Frage sorgfältig durch und beantworte sie so genau wie möglich.
- Pro Frage ist nur ein Kreuz als Antwort erlaubt.
- Wenn du dich vertan hast, male bitte das Kästchen ganz schwarz aus: ■. Kreuze dann ein anderes an.
- Wenn du ein Kästchen ausgemalt hast und es danach doch ankreuzen möchtest, setze ein Kreuz unmittelbar daneben: x■
- Beantworte bitte alle Fragen.

Vielen Dank, dass du wieder mitmachst!

Tobias Ludwig Burkhard Priemer

Tobias Ludwig und Prof. Burkhard Priemer

1. Damit wir diesen Fragebogen mit dem ersten Fragebogen, den du vor den Sommerferien bearbeitet hast, verbinden können, benötigen wir wieder dein Pseudonym:

   Wie lautet der zweite Buchstabe deines Vornamens? ☐
   (Beispiel: Du heißt „Peter“, dann lautet der zweite Buchstabe „E“)

   Wie lautet der erste Buchstabe des Vornamens deiner Mutter? ☐
   (Beispiel: Deine Mutter heißt „Anna“, dann lautet der erste Buchstabe „A“)

   Wie lauten die ersten zwei Ziffern des Geburtstages deiner Mutter? ☐☐
   (Beispiel: Deine Mutter hat am 02. Mai Geburtstag, dann lauten die ersten beiden Ziffern „02“. Wenn du nicht sicher bist, lasse die Felder frei.)

   Wie lauten die ersten zwei Buchstaben der Straße in der du wohnst? ☐☐
   (Beispiel: Du wohnst in der „Frankfurter Allee“, dann lauten die ersten zwei Buchstaben „FR“. Wenn du über die Sommerferien umgezogen bist, dann trage bitte die Straße deines ehemaligen Wohnorts ein!)

2. Welche Klassenstufe besuchst du? ☐

3. Erinnerst du dich noch an das Fadenpendel aus unserer ersten Befragung von vor den Sommerferien? Hier noch einmal die Erklärung dazu:

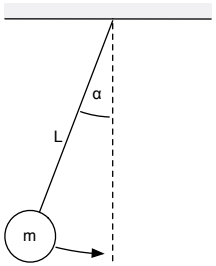
Ein Fadenpendel besteht aus einer Masse, die an einem frei beweglichen Faden aufgehängt ist. Ein Fadenpendel kann durch die folgenden Größen beschrieben werden:

T Schwingungsdauer: die Zeit, die ein Pendel benötigt, um eine komplette Schwingung durchzuführen, z. B. von der Ruhelage (Seil hängt senkrecht nach unten) zunächst ganz nach rechts, dann ganz nach links und wieder zurück zur Ruhelage. Je größer also die Schwingungsdauer, desto länger benötigt das Pendel für eine komplette Schwingung.

L Länge des Fadens, an dem die Masse aufgehängt ist

m Die angehangene Masse

α Auslenkung: der Winkel, um den das Pendel zu Beginn der Schwingung aus der Ruhelage gehoben wird



---

4. Vor den Sommerferien hast du zunächst eine Vermutung zum Zusammenhang von Pendelmasse und Schwingungsdauer aufgestellt und dann mit einem Experiment überprüft. Mit welchem Experiment zum Fadenpendel hast du vor den Sommerferien gearbeitet?

Realexperiment ☐
Computerexperiment ☐

---

5. Wie lautet nun deine Vermutung zum Zusammenhang von Pendelmasse und Schwingungsdauer?

Vergrößern der Pendelmasse bewirkt, dass die Schwingungsdauer, also die Zeit für eine komplette Schwingung, …

…größer wird. ☐
…gleich bleibt. ☐
…kleiner wird. ☐

---

6. Wie sicher bist du dir beim Aufstellen dieser Vermutung?

sehr unsicher	etwas unsicher	relativ sicher	sehr sicher
☐	☐	☐	☐

Vielen Dank noch einmal für deine Unterstützung!

C.4 ÜBERSICHT ÜBER FEHLENDE WERTE

Tabelle 29: Anzahl der fehlenden Werte pro Item

Variable	Fehlende Werte in %
wdn.handlung.1	1.39
wdn.handlung.2	0.21
wdn.handlung.3	0.21
wdn.handlung.4	0.21
wdn.handlung.5	0.11
wdn.pers.1	0.21
wdn.pers.2	0.43
wdn.pers.3	0.21
wdn.pers.4	0.43
wdn.pers.5	0.21
nfc1	0.00
nfc2	0.21
nfc3	0.53
nfc4	0.64
nfc5	0.43
nfc6	0.43
nfc7	0.85
nfc8	0.21
nfc9	0.43
nfc10	0.43
nfc11	0.11
nfc12	1.07
nfc13	0.43
nfc14	0.64
nfc15	0.21
nfc16	0.11
K12P01	0.53
K02U01	0.75
K06F02	1.49
K09P01	12.26
K10P01	2.35
K12L01	0.85
K06P04	0.96
K06L01	6.61
K07P01	10.45
K13L01	5.44
K01P01	4.05
K06P03	0.64
K06P02	2.99
B03F01	1.39
B02M01	5.65
B02F03	3.30
D04F01	1.17
D02P01	2.13
D01P02	3.30
E06M01	0.75
E04P01	6.40
E05L01	8.64
si.auf.1	0.43
si.auf.2	0.43
si.auf.3	0.43

Tabelle 29: Anzahl der fehlenden Werte pro Item

Variable	Fehlende Werte in %
si.emo.1	0.21
si.emo.2	0.21
si.emo.3	0.53
si.epi.1	0.64
si.epi.2	0.64
si.epi.3	0.53
si.epi.4	0.21
si.wert.1	0.32
si.wert.2	0.43
si.wert.3	0.32
i1.03	0.96
i1.09	0.53
i1.16	0.53
i1.18	0.75
i1.21	0.11
i6.05	1.07
i6.08	0.21
i6.11	0.96
i6.15	0.64
i6.16	1.07
i7b.04	0.43
i7b.08	0.75
i7b.16	0.85
i7b.17	0.75
i7b.22	1.39
i8.05	0.53
i8.09	1.28
i8.11	1.17
i8.14	0.32
i8.18	1.07
Hyp.follow.up	22.49

C.5 DESKRIPTIVSTATISTIK UND HISTOGRAMME DER SKALEN

Der folgende Abschnitt berichtet die deskriptivstatischen Maße Median, arithmetisches Mittel, Standardabweichung und die interne Konsistenz nach Cronbach auf Skalenebene. Zudem sind Histogramme dargestellt. Dieser Abschnitt dient in erster Linie dazu, die Rohdaten augenscheinlich zu beurteilen. Es erfolgt keine Bewertung der psychometrischen Qualität der Instrumente bzw. eine inferenzstatistische Analyse im Hinblick auf die Forschungsfragen, da diese im Rahmen von Strukturgleichungsmodellierung untersucht werden (vgl. Kapitel 5). Weiterhin sind Histogramme der Skalen aufgeführt.

C.5.1 *Argumentkategorien*

Tabelle 30: Deskriptivstatistische Maße der Skalen zur Erfassung der Stärke der Verwendung der Argumentkategorien

Subskala	Median	m	SD	α
Intuition	2.6	2.60	0.92	.87
Expertenwissen	3.6	3.56	0.89	.79
Messunsicherheiten (explizit)	3.2	3.08	0.84	.75
Daten als Evidenz	4.2	3.96	0.81	.85

Hinweise: m ... arithmetisches Mittel; SD ... Standardabweichung; α ... Interne Konsistenz nach Cronbach

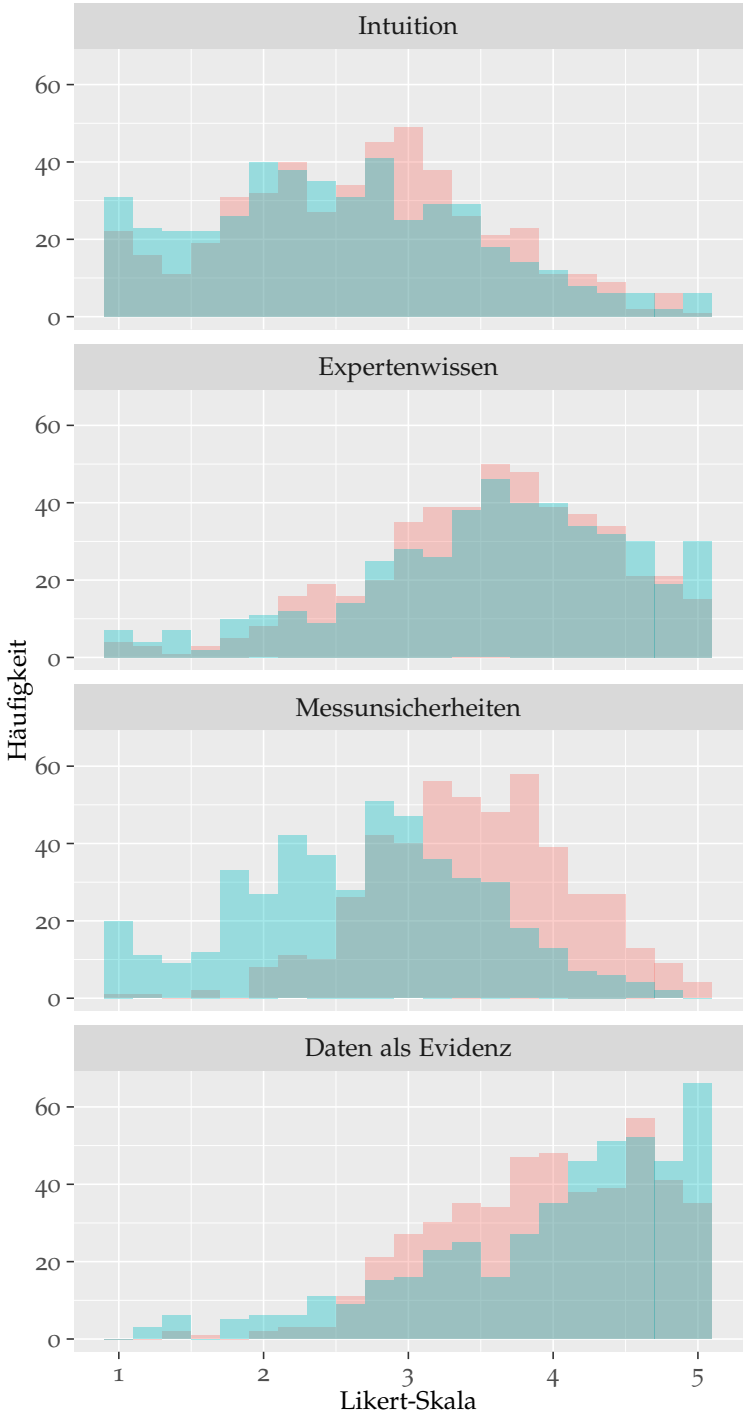


Abbildung 9: Histogramme der Subskalen des Argumentationstests. Entsprechend zu Forschungsfrage 3 zeigen die Histogramme die Verteilungen für beide Gruppen. ... Realexperiment; ... Computersimulation

C.5.2 *Kognitionsbedürfnis*

Tabelle 31: Deskriptivstatistische Maße der Skala Kognitionsbedürfnis. Da die CFA in Abschnitt 4.9 gezeigt hat, dass die negativ formulierten Items zu einem Methodenfaktor führen, und daher in der Folge ausgeschlossen wurden, berichtet diese Tabelle nur Maße für die sechs nicht-invertierten Items.

Skala	Median	m	SD	α
Kognitionsbedürfnis	3.0	2.96	0.71	.78

Hinweise: m ... arithmetisches Mittel; SD ... Standardabweichung; α ... Interne Konsistenz nach Cronbach

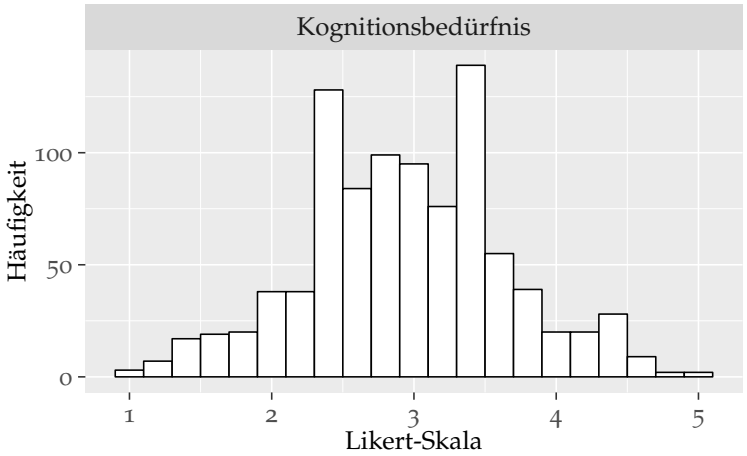


Abbildung 10: Histogramm des Mittelwerts der aus den nicht-invertierten Items bestehenden Skala zur Erfassung des Kognitionsbedürfnisses.

C.5.3 *Situationales Interesse*

Tabelle 32: Deskriptivstatistische Maße der Subskalen des Situationalen Interesses

Skala	Median	*m*	*SD*	α
Aufmerksamkeitskomponente	3.0	3.06	0.86	.78
Emotionale Komponente	3.0	2.99	0.95	.86
Epistemische Komponente	2.7	2.74	0.92	.83
Wertbezogene Komponente	2.3	2.56	0.89	.83

Hinweise: *m* ... arithmetisches Mittel; *SD* ... Standardabweichung; α ... Interne Konsistenz nach Cronbach

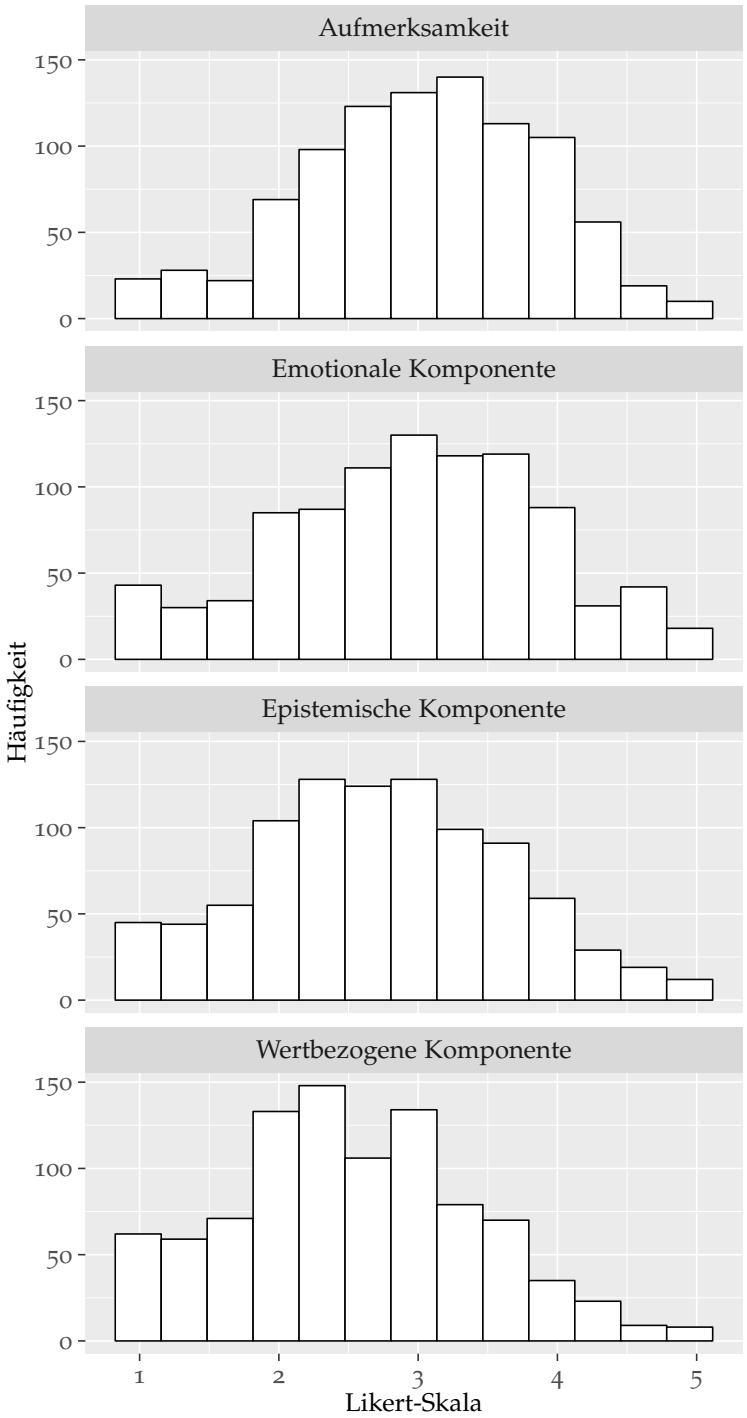


Abbildung 11: Histogramme der Subskalen des Situationalen Interesses

C.5.4 *Werteinschätzung der Naturwissenschaften*

Tabelle 33: Deskriptivstatistische Maße der Subskalen Handlungsbezogener und Persönlicher Wert der Naturwissenschaften

Skala	Median	*m*	*SD*	α
Handlungsbezogene Werteinschätzung	3.0	3.09	0.69	.59
Persönliche Werteinschätzung	2.8	2.69	0.78	.70

Hinweise: m ... arithmetisches Mittel; *SD* ... Standardabweichung; α ... Interne Konsistenz nach Cronbach

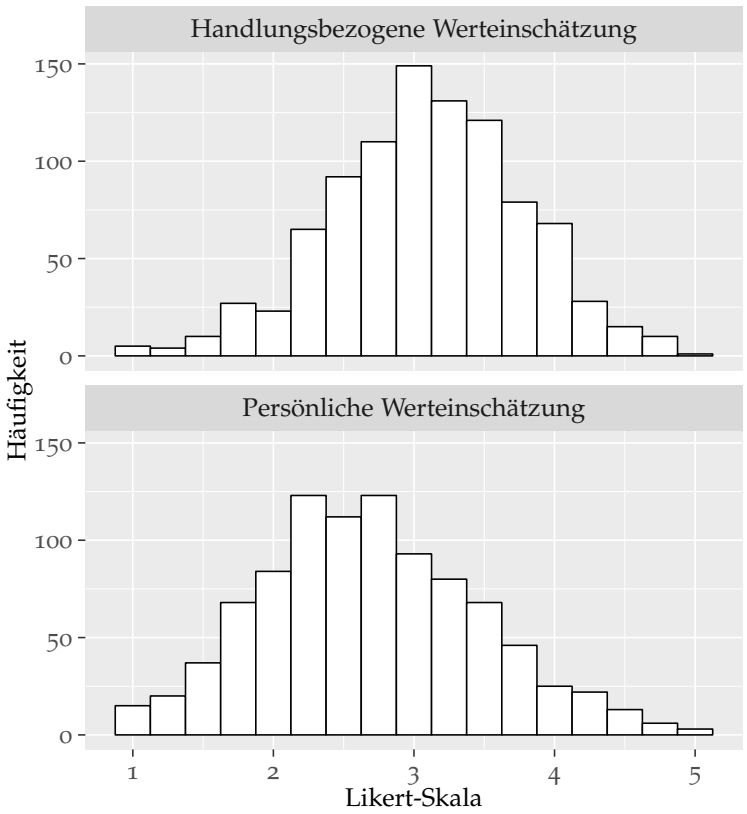


Abbildung 12: Histogramme der Subskalen Handlungsbezogener und Persönlicher Wert der Naturwissenschaften

C.6 ANALYSE DES FACHWISSENTESTS MECHANIK MIT DEM RASCH-MODELL

Tabelle 34: Itemparameter des Rasch-Modells zur Analyse des Fachwissentests Mechanik. Anmerkung: Für einige Items ist der t-Test für Infit bzw. Outfit signifikant. Da aber In- bzw. Outfit-Werte innerhalb der definierten Grenzen liegen, kann dies ignoriert werden (siehe Boone, Staver & Yale, 2014, S. 166).

Item	ξ	SE_ξ	Outfit	Outfit_t	Infit	Infit_t
K12P01	−2.13	0.10	0.99	−0.04	1.00	−0.01
K02U01	−1.36	0.08	1.02	0.43	1.02	0.41
K06F02	−0.99	0.08	0.96	−0.84	0.98	−0.57
K09P01	−0.82	0.08	1.05	1.19	1.03	1.13
K10P01	−0.42	0.07	1.05	1.89	1.03	1.53
K12L01	−0.80	0.07	0.92	−2.32	0.95	−1.92
K06P04	0.03	0.07	1.00	0.20	1.00	0.21
K06L01	−0.02	0.07	1.06	2.96	1.06	3.01
K07P01	0.67	0.08	1.05	1.55	1.04	1.28
K13L01	0.78	0.07	1.03	0.71	1.02	0.58
K01P01	2.88	0.14	1.20	1.32	1.01	0.12
K06P03	1.01	0.08	0.93	−1.78	0.95	−1.57
K06P02	0.36	0.07	0.93	−2.77	0.94	−2.83
B03F01	−1.99	0.10	0.91	−1.07	0.97	−0.41
B02M01	−0.78	0.07	0.99	−0.37	0.99	−0.34
B02F03	0.24	0.07	0.95	−2.24	0.96	−2.31
D04F01	−3.13	0.15	1.06	0.40	1.00	0.03
D02P01	−1.02	0.08	0.92	−1.78	0.96	−1.32
D01P02	−0.60	0.07	1.01	0.33	1.01	0.42
E06M01	−1.35	0.08	1.02	0.32	1.00	0.10
E04P01	−0.31	0.07	1.07	2.78	1.06	2.78
E05L01	0.14	0.07	1.03	1.55	1.03	1.52

Hinweise: ξ . . . Itemschwierigkeit, SE_ξ . . . Standardfehler der Itemschwierigkeit, Outfit . . . Itemfitparameter MNSQ Outfit, Outfit_t . . . t-Wert für Outfit-Statistik, Infit . . . Itemparameter MNSQ Infit, Infit_t . . . t-Wert für Outfit-Statistik

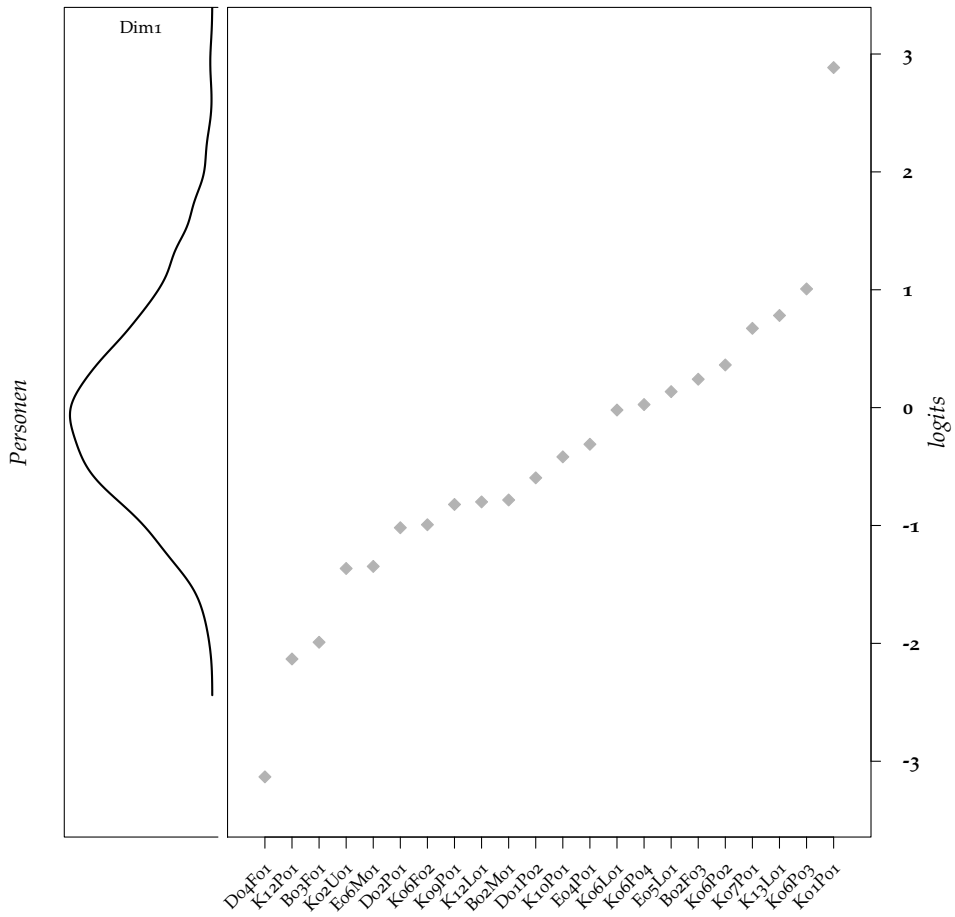


Abbildung 13: Wright-Map (Gegenüberstellung von Personenfähigkeiten und Aufgabenschwierigkeiten) für den Fachwissentest Mechanik

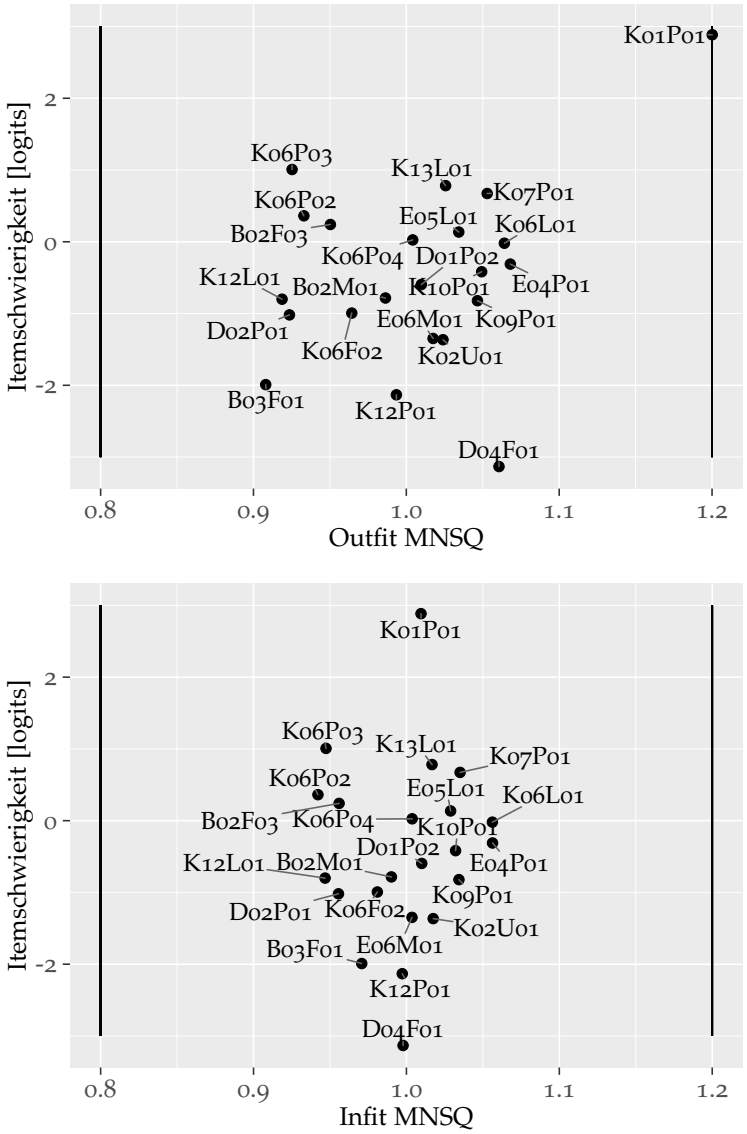


Abbildung 14: Darstellung der Itemschwierigkeit in Abhängigkeit des Item-Outfits (oben) bzw. Item-Infits für den Fachwissentest Mechanik. Beide Parameter liegen für alle Items innerhalb des akzeptierten Bereichs von 0.8 bis 1.2. Lediglich das Item `K01P01`, das auch das schwerste Item des Tests ist (Masse eines Gegenstandes auf dem Mond) schneidet den Cut-off-Wert für den Outfit, der zugehörige t-Wert ist jedoch nicht signifikant.

D

ANHANG ERGEBNISSE HAUPTSTUDIE

D.1 VERGLEICH DER MITTELWERTE IN ABHÄNGIGKEIT VON DER RICHTIGKEIT DER AUFGESTELLTEN HYPOTHESE

In Abschnitt 5.5 wird der Einfluss der Stärke der Verwendung bestimmter Argumentkategorien auf die Richtigkeit der Hypothese nach dem Experiment analysiert. Anhand der aufgestellten Hypothese wurde die Stichprobe in zwei Gruppen geteilt, um Mittelwertsdifferenzen in der Stärke der Verwendung der Argumentkategorien zu untersuchen. Dieses Verfahren weist Schwächen auf (vgl. Abschnitt 5.5) und soll hier daher nur zur Orientierung dargestellt sein. Ausgehend von dem Messmodell des 4-faktoriellen Argumentationstests wurde ein Multigruppen-konfirmatorische Faktorenanalyse (MG-CFA)-Modell mit latenter Mittelwertstruktur geschätzt. Als Referenzgruppe dient hier die Gruppe der Probanden, die nach dem Experiment eine fachlich adäquate Hypothese aufgestellt haben, entsprechend wurden die Mittelwerte in dieser Gruppe zu null fixiert (zur Referenzgruppen-Methode siehe Abschnitt 5.4). Das Modell weist eine akzeptable Passung zu den Daten auf (Tabelle 36). Die in diesem Modell geschätzten latenten Mittelwerte und Faktorvarianzen sind in Tabelle 35 dargestellt. Die Stärke der Verwendung der Argumentkategorien unterscheidet sich zwischen beiden Gruppen (richtige vs. falsche Hypothese) für alle vier Variablen signifikant. Dabei zeigen sich folgende Effekte: Zum einen ist die Verwendung der Argumentkategorie Intuition mit einer Differenz von 0.27 Einheiten auf der Likert-Skala häufiger in der Gruppe der Probanden verwendet worden, die im Anschluss an das Experiment eine falsche Hypothese aufstellen. Mit $d = 0.37$ Standardabweichungen ist von einem kleinen Effekt zu sprechen. Ein etwas größerer Effekt gleicher Richtung zeigt sich in der Verwendung der Argumentkategorie Messunsicherheiten (explizit) ($d = 0.46$). Dem gegenüber steht eine signifikant geringere Verwendung der Argumentkategorien Expertenwissen ($d = -0.39$) sowie Daten als Evidenz ($d = -0.83$).

Tabelle 35: ML-Schätzer für die Mittelwertstruktur des Argumentationstests für die Probanden mit richtiger (links) bzw. falscher Hypothese nach dem Experimentieren

	richtig ($n = 424$)			falsch ($n = 390$)			
Parameter	Unst.	*SE*	Std.	Unst.	*SE*	Std.	p
Intuition							
Varianz	0.65	0.06	1.00	0.54	0.05	1.00	
Mittelwert	0.00		0.00	0.27	0.06	0.37	$< .001$
Expertenwissen							
Varianz	0.65	0.08	1.00	0.74	0.09	1.00	
Mittelwert	0.00		0.00	-0.33	0.07	-0.39	$< .001$
Messunsicherheiten							
Varianz	0.31	0.06	1.00	0.22	0.06	1.00	
Mittelwert	0.00		0.00	0.21	0.05	0.46	$< .001$
Daten als Evidenz							
Varianz	0.40	0.05	1.00	0.60	0.07	1.00	
Mittelwert	0.00		0.00	-0.64	0.06	-0.83	$< .001$

Hinweise: Unst.. . . unstandardisierter Parameter; Std.. . . Standardisierter Parameter; *SE*. . . Standardfehler, p. . . Irrtumswahrscheinlichkeit (Welch-James t-Statistik). Standardisierte Mittelwertparameter können wie die d-Statistik nach Cohen (1988) interpretiert werden (vgl. Abschnitt E.2).

Tabelle 36: Fit-Indizes für das Modell `hyp.richtig`

	χ^2	df	p	χ^2/df	CFI	RMSEA [95 % C.I.]; p	SRMR
`hyp.richtig`	2276.8	1418	$< .01$	1.61	.92	.04 [.036; .041]; 1	.06

E

ANMERKUNGEN ZU VERWENDETEN STATISTISCHEN VERFAHREN

E.1 KRITERIEN FÜR DIE BEURTEILUNG DER MODELLANPASSUNGSGÜTE BEI KONFIRMATORISCHEN FAKTORENANALYSEN UND STRUKTURGLEICHUNGSMODELLEN

Der folgende Abschnitt erläutert die in der vorliegenden Arbeit herangezogenen Kriterien zur Beurteilung der Modellgüte im Rahmen der konfirmatorischen Faktorenanalysen und der Anpassung von Strukturgleichungsmodellen.

E.1.1 *Inferentielle und deskriptive Kriterien der Modellanpassungsgüte*

Bei der Beurteilung der Modellgüte sind prinzipiell zwei Vorgehensweisen zu unterscheiden: Zum einen *inferentiell* über den χ^2-Test, der auf der gewichteten Abweichung von beobachteter ($\boldsymbol{S}$) und modellimplizierter Kovarianzmatrix ($\boldsymbol{\Sigma}$) unter Berücksichtigung der Stichprobengröße beruht. Zum anderen ein *deskriptives* Vorgehen durch die Beurteilung verschiedener Fit-Indizes (CFI, RMSEA, SRMR) (Moosbrugger & Kevala, 2008). Der χ^2-Test evaluiert die Nullhypothese

$$H_0 : \boldsymbol{S} = \boldsymbol{\Sigma}, \tag{24}$$

d. h. die Fähigkeit des Modells, die Stichprobenkovarianzmatrix exakt zu reproduzieren. In jüngerer Zeit ist eine Kontroverse über die Anwendung dieser Fit-Indizes entbrannt. Die Gefahr der Fit-Indizes liegt in der Überinterpretation eines im Sinne des χ^2-Tests missspezifizierten Modells, d. h. die Ablehnung der Nullhypothese $\boldsymbol{S} = \boldsymbol{\Sigma}$ (für eine Übersicht siehe z. B. Kenny, 2014). Dem gegenüber stehen eine Reihe von Problemen bei der alleinigen Betrachtung des χ^2-Tests:

1. Der χ^2-Test verhält sich bei großen Stichproben sehr sensitiv und kann schon bei geringen Abweichungen zwischen beobachteter und implizierter Kovarianzmatrix signifikant werden (Brown, 2006; Kline, 2016).

2. Der χ^2-Test testet die äußerst restriktive Hypothese $\boldsymbol{S} = \boldsymbol{\Sigma}$, die oftmals nicht besonders praktikabel ist (Brown, 2006; Kline, 2016)

3. Die Power des χ^2-Tests ist bei geringen Stichprobengrößen nicht ausreichend, so dass Fehlspezifikationen ggf. nicht entdeckt werden können. Weiterhin ist zu berücksichtigen, dass außerdem die Komplexität des Modells, z. B. ausgedrückt in der Zahl der

Freiheitsgrade df, ebenfalls eine Rolle spielt: MacCallum et al. (1996) empfehlen, dass insbesondere bei Modellen mit einer hohen Zahl an Freiheitsgraden und einer geringen Stichprobengröße ein exakter Fit nicht möglich ist. Beziehungsweise reziprok formuliert: Bei einer hohen Zahl an Freiheitsgraden und einem schlechten Modellfit liegt schon bei einer geringen Stichprobenzahl eine ausreichend hohe Power vor (Beispiel: $df = 100$, Power = .8, ergibt eine notwendige Stichprobengröße von $n = 132$, siehe MacCallum et al., 1996, S. 144).

In der vorliegenden Arbeit wird, der χ^2-Test daher nicht interpretiert. Der Fokus in der Beurteilung der verschiedenen Gütemaße liegt auf den deskriptiven Fit-Indizes. Folgende Kriterien wurden zur Interpretation der Modellgütemaße herangezogen (alle Cut-off-Werte beziehen sich auf die Verwendung von robusten ML-Schätzverfahren, siehe Brown, 2006):

χ^2-TEST Für einen exakten Modellfit soll der χ^2-Test nicht signifikant sein, $p > .05$. Als „Daumenregel" soll das Verhältnis $\chi^2/df < 3$ gelten. (Moosbrugger & Kevala, 2008), wobei auch dieser Wert nicht überinterpretiert werden sollte, da kein anerkannter Cut-off-Wert existiert (Brown, 2006; Hu & Bentler, 1999).

CFI Der Comparative-Fit-Index (CFI) evaluiert die Güte des spezifizierten Modells in Relation zu einem geschachtelten, aber beschränkterem Null-Modell. In diesem Vergleichsmodell wird keinerlei Kovarianz zwischen den Indikatoren zugelassen, der Fit des spezifizierten Modells wird also mit der Annahme verglichen, dass keinerlei Beziehungen zwischen den Indikatoren existieren. Wenngleich dieses Kriterium relativ großzügig erscheint, wird dem CFI im Allgemeinen eine hohe Bedeutung bei der Beurteilung der Anpassungsgüte zugesprochen. Der CFI kann Werte zwischen 0 und 1 annehmen, wobei Werte nahe eins einem guten Modellfit entsprechen (Brown, 2006). Für einen guten Modellfit sollte der CFI um .95 oder darüber liegen (Hu & Bentler, 1999). Der CFI ist nicht von der Stichprobengröße beeinflusst (Schermelleh-Engel, Moosbrugger & Müller, 2003). Als Cut-off-Wert für den CFI werden in der Literatur verschiedene Größen genannt. Bei Werten CFI$<$ .90 sollten die spezifizierten Modelle abgelehnt werden. Werte von .90 bis .95 stehen für akzeptablen Modellfit (Brown, 2006; West, Aaron & Wu, 2012). Brown (2006) empfiehlt entsprechend der neueren Studien von Hu und Bentler (1999) für einen guten Modellfit einen strengeren Cut-off in der Nähe von .95 oder darüber („close to .95 or greater", S. 87), betont jedoch gleichzeitig die Bedeutung einer gleichzeitigen Interpretation von mehreren Fit-Indizes. In der vorliegenden Arbeit werden CFI-Werte $>$ .90 akzeptiert.

RMSEA Der RMSEA ist ein populationsbasierter Index, der den Fehler des Modellfits bestimmt (error of approximation) und ist daher konzeptuell anders zu verstehen als der χ^2-Test. Der RMSEA bevorzugt dabei Modelle mit einer geringeren Anzahl an frei geschätzten Parametern (bei gleicher absoluter Modellgüte), bestraft also die Modellkomplexität. Der RMSEA ist unbeeinflusst von der Stichprobengröße (Brown, 2006). RMSEA-Werte von 0 stehen für einen perfekten Fit, RMSEA < .06 für einen guten Fit (Hu & Bentler, 1999), RMSEA$<$ .08 für einen akzeptablen Fit, Modelle mit RMSEA$>$ .1 sollten abgelehnt werden (Brown, 2006). Für den RMSEA kann ein Konfidenzintervall (typischerweise 90 %) bestimmt werden. Dieses gibt Aufschluss über die Genauigkeit des erhaltenen Gütemaßes. Für einen akzeptablen Modellfit sollte die obere Grenze des Konfidenzintervalls entsprechend den Wert .08 nicht übersteigen. Ferner können Abweichungen der Hypothese H_0 :RMSEA $\leq$ .05 auf Signifikanz getestet werden. Nichtsignifikante Abweichungen sind als guter Modellfit zu interpretieren (Brown, 2006).

SRMR Der Standardized Root Mean Square Residual (SRMR) untersucht die Passung des Modells auf absoluter Ebene indem die durchschnittliche Abweichung zwischen den beobachteten Korrelationen und den modellimplizierten Korrelationen bestimmt wird. Ein für akzeptable bis gute Modellfite allgemein akzeptierter Cut-off-Wert für den SRMR ist $<$ .08 (Brown, 2006).

Für weiterführende Darstellungen siehe auch Brown (2006), Hu und Bentler (1999), Kenny (2014), MacCallum et al. (1996), Schermelleh-Engel et al. (2003). Eine tabellarische Übersicht einer Vielzahl von Gütemaßen findet sich in West et al. (2012, S. 212).

E.1.2 *Kriterien für den Vergleich hierarchisch geschachtelter Modelle*

In der vorliegenden Arbeit werden sogenannte *nested models* oder hierarchisch geschachtelte Modelle verglichen. „Mehrere CFA-Modelle werden als hierarchisch geschachtelt bezeichnet, wenn sie dieselbe Modellstruktur aufweisen, sich jedoch in der Anzahl der fixierten oder freigesetzten Parameter unterscheiden. Sie heißen hierarchisch geschachtelt, weil in den verschiedenen Modellen zunehmend mehr Parameter fixiert oder freigesetzt werden, so dass sie durch Parameterrestriktionen bzw. Freisetzung ineinander überführt werden können, während die Modellstruktur ansonsten erhalten bleibt“ (Moosbrugger & Kevala, 2008, S. 420). In dieser Arbeit werden z. B. Modelle mit Messinvarianzbedingungen miteinander verglichen.

Ob sich ein restriktiveres Modell im Vergleich zu einem anderen, weniger restriktiven Modell unterscheidet, wird oftmals über einen χ^2-*likelihood-ratio*-Test (LR-Test oder auch $\Delta\chi^2$-Differenzen-Test) auf

Signifikanz geprüft (Kline, 2016). Dieser Test weist jedoch die gleichen Schwächen wie der χ^2-Test zur Evaluation des absoluten Fits auf (vgl. Abschnitt E.1): Die Power des χ^2-Differenzen-Tests ist ebenfalls stark von der Stichprobengröße abhängig. Bei kleinen Stichprobengrößen (wie hier z. B. in der Testentwicklungsstudie, vgl. Anhang B) kann der LR-Test leicht nicht-signifikante Ergebnisse aufweisen, und daher evtl. vorhandene bedeutsame Unterschiede zwischen den geschachtelten Modellen aufgrund geringer Power nicht aufklären. Dies ist insbesondere bei der Evaluation der Messinvarianzbedingungen problematisch, da sich dadurch die Wahrscheinlichkeit für einen β-Fehler, d. h. Beibehalten der H_0 bei Vorliegen der H_1 in der Population erhöht (West et al., 2012). Im Gegensatz dazu kann der χ^2-Differenzentest bei sehr großen Stichproben, wie sie z. B. in der vorliegenden Arbeit in der Hauptuntersuchung vorliegen, leicht Signifikanz erreichen, obwohl die Differenz (im Sinne einer Effektstärke) nur marginal und von geringer inhaltlicher Bedeutung ist (Kline, 2016, S. 399).

Eine Reihe von Forschungsarbeiten kritisieren daher die Verwendung des χ^2-Differenzen-Tests zur Beurteilung hierarchischer Modelle, insbesondere dann, wenn zur Beurteilung des übergreifenden Modellfits alternative Fit-Indizes wie RMSEA, SRMR und CFI herangezogen worden sind, z. B. weil der Test gegen die Hypothese eines exakten Modellfits durch den χ^2-Test als zu streng bzw. inpraktikabel abgetan wird (für eine Übersicht siehe z. B. Cheung & Rensvold, 2002; Meade, Johnson & Braddy, 2008). In der Folge wurden eine Reihe alternativer Kriterien vorgeschlagen. Cheung und Rensvold (2002) haben die Eignung von Differenzen der CFI-Werte geschachtelter Modelle in Simulationsstudien untersucht und stellten fest, dass ΔCFI als stichprobenunabhängiges Maß geeignet ist, Modellunterschiede aufzudecken. Der ΔCFI-Wert kann dabei positive oder negative Werte annehmen. Wenn die Nullhypothese (z. B. „Es liegt Messinvarianz vor“) wahr ist, kann der ΔCFI, bedingt durch die Abhängigkeit des CFI von der Zahl der Freiheitsgrade df, größer werden. Wird der ΔCFI-Wert negativ, spricht das für das für eine Verletzung der Invarianzbedingung. Dabei werden verschiedene Cut-off-Werte berichtet: Bei $\Delta\text{CFI} < -.01$ soll nach Cheung und Rensvold (2002) die Nullhypothese verworfen werden. Meade et al. (2008) empfehlen ebenfalls die Verwendung von ΔCFI und schlagen einen strengeren Cut-off-Wert von $\Delta\text{CFI} \leq -.002$ vor, F. F. Chen (2007) schlägt bei Stichproben mit $n \leq 300$ einen Cut-off-Wert von $\Delta\text{CFI} \leq -.005$ vor, ergänzt durch $\Delta\text{RMSEA} \geq .010$. Bei Stichproben mit $n > 300$ empfiehlt F. F. Chen (2007, S. 501) $\Delta\text{CFI} \leq -.010$ und $\Delta\text{RMSEA} \geq .015$ als Cut-off-Kriterien.

In der vorliegenden Arbeit werden beim Vergleich hierarchisch geschachtelter Modelle sowohl der χ^2-Differenzentest als auch die Differenz in den approximativen Fit-Indizes interpretiert. Zeigt also der χ^2-Differenzentest eine signifikante Verschlechterung des Modell-Fits,

wird hier anhand der Cut-off-Kriterien für ΔCFI und ΔRMSEA die empirische Bedeutsamkeit der Invarianz geprüft.

E.2 EFFEKTGRÖSSEN

E.2.1 *Effektstärke r für Zusammenhänge*

Zur Untersuchung von Zusammenhangshypothesen wurden in den Strukturgleichungsmodellen standardisierte Pfadkoeffizienten berechnet, die wie das Effektstärkemaß r interpretiert werden können. Dieses skaliert von -1.0 bis $+1.0$, wobei durch das Vorzeichen die Richtung des Effekts angegeben wird. Zur Einschätzung der Effektstärke werden die bekannten Intervalle nach Cohen (1988) herangezogen:

$|r| \geq .1$ ... kleiner Effekt
$|r| \geq .3$ ... mittlerer Effekt
$|r| \geq .5$ ... großer Effekt

Durch das Quadrat dieses Maßes r^2 kann der Anteil der erklärten Varianz bestimmt werden. Ein Effekt von $r = .3$ entspricht daher einem Anteil aufgeklärter Varianz von $r^2 = 9\,\%$. Einige Autoren warnen davor, dass bei quadrierten Zusammenhangsmaßen die Gefahr besteht, substantielle Effekte zu unterschätzen. Nach Kline (2013, S. 128) beträgt der Anteil aufgeklärter Varianz in einem univariaten Fall meist nicht mehr als 10 %. Cohen (1988, S. 78) weist ebenfalls auf diese Problematik hin und betont, dass daher der Anteil aufgeklärter Varianz in den Sozialwissenschaften nicht mit den in den Naturwissenschaften erreichten Beträgen verglichen werden darf.

E.2.2 *Effektstärke ϕ für 2x2-Kontingenzanalysen*

In der vorliegenden Arbeit wird die Un- bzw. Abhängigkeit von Gruppenzugehörigkeit und Hypothesenwahl, zwei binär nominalskalierter Variablen, über 2x2-Kontingenzanalysen bestimmt. Als Maß für die Effektstärke kann der Korrelationskoeffizient Cramers ϕ herangezogen werden, der sich im Fall von 2x2-Tabellen aus dem χ^2-Wert und der Fallzahl n zu

$$\phi = \sqrt{\frac{\chi}{n}} \tag{25}$$

ergibt (Cohen, 1988, S. 223). Alternativ kann ϕ über den Korrelationskoeffizienten nach Pearson r bestimmt werden. Die Interpretation von ϕ verläuft im Falle von 2x2-Tabellen analog zu r (siehe oben).

E.2.3 *Effektgrößen für Mittelwertsunterschiede - Cohens d*

Im Rahmen der Gruppenvergleiche wurden in der vorliegenden Arbeit standardisierte Mittelwertsunterschiede berechnet, die wie die *d*-

Statistik nach Cohen (1988) interpretiert werden können. Dabei wird der Unterschied in den Stichproben-Mittelwerten durch die beobachtete Standardabweichung $\overline{\theta}$ dividiert, d. h. $\overline{d} = \overline{\mu}_1 - \overline{\mu}_2/\overline{\theta}$, um einen Schätzer für die Effektstärke d in der Population zu erhalten. In Abhängigkeit von der Richtung der Differenzbildung kann d positive oder negative Werte annehmen.

Bei der Standardisierung des Mittelwertsunterschieds muss jedoch berücksichtigt werden, dass in Multigruppen-Strukturgleichungsmodellen (**MG-SEM!**) die Faktorvarianzen ($\overline{\theta}^2$) in beiden Teilstichproben unterschiedlich sein *können*, da im Gegensatz zu klassischen Varianzanalysen die Varianzhomogenität (Homoskedastizität) hier keine notwendige Bedingung darstellt (Green & Thompson, 2012, S. 408). Daher gibt es, in Abhängigkeit der verwendeten Varianz bzw. Standardabweichung, konzeptuell verschiedene Möglichkeiten, um Mittelwertsunterschiede zu standardisieren. In der vorliegenden Arbeit sind die Faktorvarianzen in beiden Gruppen meist von vergleichbarer Größe, so dass der zu erwartende Unterschied zwischen verschiedenen Arten der Standardisierung vernachlässigbar klein ist. In den beschriebenen Situationen, in denen die Gruppenvarianzen nicht vergleichbar sind, kann die gepoolte Varianz $\overline{\theta}^2_{\text{pooled}}$,

$$\overline{\theta}_{\text{pooled}} = \sqrt{\frac{(n_1 - 1)\overline{\theta}_1^2 + (n_2 - 1)\overline{\theta}_2^2}{n_1 + n_2 - 2}}, \tag{26}$$

in der Standardisierung verwendet werden, wobei n_1 und n_2 die Stichprobenumfänge je Gruppe und $\overline{\theta}_1^2$ und $\overline{\theta}_2^2$ die Stichprobenvarianzen darstellen (z. B. Hancock, 2001, S. 374). Da die Gruppenvarianzen in der vorliegenden Arbeit überwiegend vergleichbar sind, wurde darauf jedoch verzichtet und nur die ungepoolte Varianz zur Standardisierung herangezogen. Dieses Vorgehen ist gerechtfertigt, wie das folgende Beispiel skizziert: In Tabelle 16 sind die Parameter für die Gruppenunterschiede ohne Kontrolle der Prädiktoren dargestellt. Zur Standardisierung der dort dargestellten Parameter wurde die latente Faktorvarianz der jeweiligen Gruppe verwendet, hier also $\theta^2_{\text{mu, V}} = 0.22$. Der so standardisierte Mittelwertsunterschied bei der Verwendung der Kategorie Messunsicherheiten (explizit) beträgt $d = 0.53/\sqrt{0.22} \approx 1.14$, zugunsten der Gruppe der Probanden, die mit dem Realexperiment gearbeitet haben. Würde die gepoolte Varianz zur Schätzung des standardisierten Mittelwertsunterschieds herangezogen werden, würde sich ein Wert von

$$d_{\text{pooled}} = \frac{\overline{\mu}_R - \overline{\mu}_V}{\sqrt{\frac{(n_R - 1)\theta_R^2 + (n_V - 1)\theta_V^2)}{n_V + n_R - 2}}} = -1.24 \tag{27}$$

ergeben. Dieser Unterschied zwischen der Verwendung der gepoolten bzw. ungepoolten Varianz ist daher als vernachlässigbar gering

einzustufen (für die anderen in Tabelle 16 berichteten signifikanten Effekte ist der Unterschied noch geringer als in diesem Beispiel).

Die Interpretation von standardisierten Mittelwertsunterschieden erfolgt bei der Verwendung von MG-SGM und latenten Mittelwerten ebenso wie in univariaten Varianzanalysen mit manifesten Variablen, wobei die Standardisierung jedoch um Messfehlervarianzen korrigiert ist: Ein standardisierter Mittelwertsunterschied von $d = 0.25$ bedeutet, dass die beiden Populationen auf dieser Variable auf dem latenten Konstrukt rund ein Viertel einer messfehlerfreien Standardabweichung auseinanderliegen. Es können daher die bekannten Richtlinien für kleine, mittlere und große Effekte herangezogen werden. Es werden hier die nach Cohen (1988) geltenden Intervalle zur Einschätzung der Effekte verwendet, wobei einige Autoren jedoch bei der Interpretation von Effektstärken in Modellen mit latenten Variablen noch Forschungsbedarf sehen (Gignac & Szodorai, 2016; Green & Thompson, 2012; Hancock, 2001):

$d \geq 0.2$ ... kleiner Effekt
$d \geq 0.5$ ... mittlerer Effekt
$d \geq 0.8$ ... großer Effekt

E.2.4 *Interpretation von odds ratios*

Zur Interpretation von *odds ratios* (OR) als Maß für eine Effektstärke in logistischen Regressionen liegen derzeit verschiedene Vorschläge vor (H. Chen, Cohen & Chen, 2010; Ferguson, 2009; C. K. Haddock, Rindskopf & Shadish, 1998). Da das Auftreten bzw. Nicht-Auftreten des Ereignisses in der vorliegenden Arbeit etwa gleichverteilt ist (vgl. Abbildung 6) werden hier zur Interpretation der *odds ratios* die Kriterien nach Olivier, May und Bell (2016) herangezogen:

$OR \geq 1.32$ ... kleiner Effekt
$OR \geq 2.38$ ... mittlerer Effekt
$OR \geq 4.70$ ... großer Effekt

Borenstein, Hedges, Higgins und Rothstein (2009, S. 47) berichten eine Möglichkeit, um *odds ratios* in Standardabweichungen d zu überführen:

$$d = \ln(OR)\frac{\sqrt{3}}{\pi} \tag{28}$$

Wird dieser Vorschlag auf die o. g. Grenzen nach Olivier et al. (2016) angewandt, ergeben sich die bekannten Kriterien für die d-Statistik: Bei $OR = 2.38$ ergibt sich nach Gl. 28 $d = \ln(2.38)\frac{\sqrt{3}}{\pi} \approx 0.48$ was ebenfalls einem mittleren Effekt nach Cohen (1988) entspricht (vgl. Abschnitt E.2.3). Beide Interpretationen sind daher zueinander konsistent.

E.3 ANMERKUNGEN ZU LOGISTISCHEN REGRESSIONSMODELLEN

In der vorliegenden Arbeit werden logistische Regressionsmodelle herangezogen, um den Einfluss der Verwendung der Argumentkategorien auf die Richtigkeit der nach dem Experiment aufgestellten Hypothesen zu untersuchen. Im folgenden Abschnitt sind einige Grundlagen zur logistischen Regression dargestellt, die zur Interpretation der Ergebnisse nötig sind. Als weiterführend seien die Werke von Behnke (2015), Hosmer et al. (2013), Urban (1993) genannt.

Logit-Modelle sind dazu geeignet, Einflüsse verschiedenster Faktoren auf eine diskrete, z. B. binäre bzw. kategoriale abhängige Variable zu untersuchen (z. B. Döring & Bortz, 2016, S. 678). Im vorliegenden Falle ist die Richtigkeit der aufgestellten Hypothese binär kodiert (siehe oben). Durch den Einsatz eines multiplen logistischen Regressionsmodells können die partiellen Effektstärken der Argumentkategorien (als unabhängige Variablen) simultan und kontrolliert geschätzt werden (Urban, 1993). Da die abhängige Variable beschränkt ist (hier zwischen 0 und 1), „ist eine lineare Funktion für eine Prognose oder Schätzung offensichtlich ungeeignet“ (Behnke, 2015, S. 16). Um diese Wahrscheinlichkeit P_i in eine echte kontinuierliche Größe zu überführen, welche auf einen unbeschränkten Wertebereich zurückgreifen kann, wird zum einen das Verhältnis aus der Wahrscheinlichkeit für das Ereignis und der entsprechenden Gegenwahrscheinlichkeit für dieses Ereignis gebildet. Dieses Verhältnis zweier Wahrscheinlichkeiten wird auch als *odds* (dt. „Gewinnchance“) bezeichnet. Die *odds* werden dann logarithmiert, um die untere Grenze des Wertebereichs für P_i aufzuheben:

$$L_i(y = 1) = \ln(\text{odds}) = \ln\left(\frac{P_i}{1 - P_i}\right). \tag{29}$$

Bei L_i handelt es sich um den natürlichen Logarithmus der *odds*, auch als „Logit“ bezeichnet. In der multivariaten logistischen Regressionsgleichung können dann die Einflüsse der n unabhängigen Prädiktor-Variablen auf die Logitvariable bestimmt werden:

$$L_i(y = 1) = \ln\left(\frac{P_i}{1 - P_i}\right) = \beta_1 \cdot X_1 + \cdots + \beta_n \cdot X_n + \alpha \tag{30}$$

In Gl. 30 beschreibt α eine Konstante, die alle anderen Einflüsse auf die abhängige Variable zusammenfasst. Der Parameter β_i gibt die Stärke und Richtung des Einflusses jeder i-ten unabhängigen Variablen – bei gleichzeitiger Konstanthaltung aller anderen unabhängigen Variablen – an (Urban, 1993). Die Parameter β_i sind jedoch nicht mehr einfach zu interpretieren, da sie den Einfluss der unabhängigen Variable auf die abhängige Variable in ihrer Logit-Form und nicht in ihren ursprünglichen Messwerten darstellen.

Zur Interpretation kann daher Gl. 30 nach P_i umgestellt werden, so dass die Wahrscheinlichkeit für $P(y = 1)$ bei einer beobachteten Ausprägung von x_i bestimmt werden kann (Hosmer et al., 2013; Urban, 1993). Der Zusammenhang zwischen der Wahrscheinlichkeit P_i und der Ausprägung der unabhängigen Variablen entspricht dann einer logistischen Funktion; einer sigmoider Kurve, die sich asymptotisch den Grenzen 0 und 1 annähert. Alternativ können durch Entlogarithmierung von Gl. 30 sog. *odds ratios* (OR, auch Chancenverhältnis) aus den β-Parametern in Form von $\exp(\beta_i)$ bestimmt werden. Bei *odds ratios* handelt es sich um Multiplikationsfaktoren, „die Veränderungen im Wahrscheinlichkeitsverhältnis der beiden Handlungsalternativen [bei Erhöhung der unabhängigen Variablen x_i um eine Skaleneinheit] angeben“ können (Urban, 1993, S. 41).

LITERATUR

Abd-El-Khalick, F. (2008). Modeling Science Classrooms After Scientific Laboratories: Recommendations for Research and Implementation. In R. A. Duschl & R. Grandy (Hrsg.), *Teaching Scientific Inquiry: Recommendations for Research and Implementation* (S. 99–117). Rotterdam: Sense Publishers.

Abrahams, I. (2017). Minds-On Practical Work for Effective Science Learning. In K. S. Taber & B. Akpan (Hrsg.), *Science Education: An International Course Companion* (S. 403–413). New Directions in Mathematics and Science Education. SensePublishers.

Albert, E. (1978). Development of the Concept of Heat in Children. *Science Education*, *62*(3), 389–399.

Alexander, P. A., Fives, H., Buehl, M. M. & Mulhern, J. (2002). Teaching as Persuasion. *Teaching and Teacher Education*, *18*(7), 795–813.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education & Joint Committee on Standards for Educational and Psychological Testing. (2014). *Standards for educational and psychological testing*.

Backhaus, U. (2000). Die schnellste Verbindung. Zwangskräfte bei krummlinigen Bewegungen. *Der mathematische und naturwissenschaftliche Unterricht*, *53*(2).

Beauducel, A. & Herzberg, P. Y. (2006). On the Performance of Maximum Likelihood Versus Means and Variance Adjusted Weighted Least Squares Estimation in CFA. *Structural Equation Modeling: A Multidisciplinary Journal*, *13*(2), 186–203.

Beaujean, A. A. (2014). *Latent Variable Modeling Using R: A Step-by-Step Guide*. New York, NY: Routledge.

Behnke, J. (2015). *Logistische Regressionsanalyse: Eine Einführung*. Wiesbaden: Springer VS.

Benoit, W. L., Hample, D. & Benoit, P. J. (1992). *Readings in Argumentation*. Berlin, New York: Walter de Gruyter.

Berland, L. K. & Lee, V. R. (2010). Anomalous Graph Data and Claim Revision During Argumentation. In *Proceedings of the 9th International Conference of the Learning Sciences - Volume 2* (S. 314–315). ICLS '10. Chicago, Illinois: International Society of the Learning Sciences.

Berland, L. K. & Reiser, B. J. (2011). Classroom Communities' Adaptations of the Practice of Scientific Argumentation. *Science Education*, *95*(2), 191–216.

Berndt, T. (2014). *Empirische Evaluation physikdidaktischer Testinstrumente an Schülerinnen und Schülern der 8. und 9. Jahrgangsstufe: Fachwissen Mechanik, Kognitionsbedürfnis, persönliche Einstellungen gegenüber Physik und einer Experimentieraufgabe* (Masterarbeit, Humboldt-Universität zu Berlin).

Betsch, C. (2004). Präferenz für Intuition und Deliberation (PID). *Zeitschrift für Differentielle und Diagnostische Psychologie*, *25*(4), 179–197.

Bless, H., Wänke, M., Bohner, G., Fellhauer, R. F. & Schwarz, N. (1994). Need for Cognition: Eine Skala zur Erfassung von Engagement und Freude bei Denkaufgaben. *Zeitschrift für Sozialpsychologie*, *25*(2), 147–154.

Bond, T. & Fox, C. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Boone, W. J., Staver, J. R. & Yale, M. S. (2014). *Rasch Analysis in the Human Sciences*. New York, NY: Springer.

Borenstein, M., Hedges, L. V., Higgins, J. P. T. & Rothstein, H. R. (2009). *Introduction to Meta Analysis*. West Sussex: Wiley.

Bors, D. A., Vigneau, F. & Lalande, F. (2006). Measuring the need for cognition: Item polarity, dimensionality, and the relation with ability. *Personality and Individual Differences*, *40*(4), 819–828.

Brandenburger, M. (2017). *Was beeinflusst den Erfolg beim Problemlösen in der Physik? Eine Untersuchung mit Studierenden*. Berlin: Logos Verlag.

Brell, C. (2008). *Lernmedien und Lernerfolg - reale und virtuelle Materialien im Physikunterricht.* Berlin: Logos Verlag.

Brell, C., Theyßen, H., Schecker, H. & Schumacher, D. (2006). Simulation, IBE, Realexperiment – Lerneffizienz durch Neue Medien. In A. Pitton (Hrsg.), *Lehren und Lernen mit neuen Medien* (S. 81–83). Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in Paderborn 2005. Münster: LIT Verlag.

Brown, T. A. (2006). *Confirmatory Factor Analysis for Applied Research*. New York: Guilford Press.

Bühner, M. (2010). *Einführung in die Test- und Fragebogenkonstruktion*. München: Pearson.

Bunge, M. A. (1962). *Intuition and Science*. Greenwood Press.

Bybee, R. W. (1997). *Achieving Scientific Literacy: From Purposes to Practices*. Portsmouth: Heinemann Educ Books.

Byrne, B. M. & Stewart, S. M. (2006). The MACS Approach to Testing for Multigroup Invariance of a Second-Order Structure: A Walk Through the Process. *Structural Equation Modeling: A Multidisciplinary Journal*, *13*(2), 287–321.

Cacioppo, J. T., Petty, R. E. & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment*, *48*(3), 306–307.

Cacioppo, J. T. & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42*(1), 116–131.

Cacioppo, J. T., Petty, R. E., Feinstein, J. & Jarvis, B. (1996). Dispositional differences in cognitive motivation: The life and times of individuals low versus high in need for cognition. *Psychological Bulletin*, *119*, 197–253.

Cappell, J. (2013). *Fachspezifische Diagnosekompetenz angehender Physiklehrkräfte in der Ersten Ausbildungsphase*. Berlin: Logos Verlag.

Chalmers, A. F. (2007). *Wege der Wissenschaft - Einführung in die Wissenschaftstheorie*. Heidelberg: Springer Verlag.

Chen, F. F. (2007). Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464–504.

Chen, H., Cohen, P. & Chen, S. (2010). How Big is a Big Odds Ratio? Interpreting the Magnitudes of Odds Ratios in Epidemiological Studies. *Communications in Statistics - Simulation and Computation*, *39*(4), 860–864.

Cheung, G. W. & Rensvold, R. B. (2002). Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255.

Chinn, C. A. & Brewer, W. F. (1993). The Role of Anomalous Data in Knowledge Acquisition: A Theoretical Framework and Implications for Science Instruction. *Review of Educational Research*, *63*(1), 1–49.

Chinn, C. A. & Brewer, W. F. (1998). An empirical test of a taxonomy of responses to anomalous data in science. *Journal of Research in Science Teaching*, *35*(6), 623–654.

Chinn, C. A., Buckland, L. A. & Samarapungavan, A. (2011). Expanding the Dimensions of Epistemic Cognition: Arguments From Philosophy and Psychology. *Educational Psychologist*, *46*(3), 141–167.

Chinn, C. A. & Malhotra, B. A. (2002). Children's Responses to Anomalous Scientific Data: How Is Conceptual Change Impeded? *Journal of Educational Psychology*, *94*(2), 327–43.

Chinn, C. A. & Samarapungavan, A. (2001). Distinguishing between Understanding and Belief. *Theory into Practice*, *40*(4), 235–41.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd). Routledge.

Cohen, J., Cohen, P., West, S. G. & Aiken, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Mahwah, N.J.: Routledge.

Copeland, E. J., Sami, M. & Tsujikawa, S. (2006). Dynamics of dark energy. *International Journal of Modern Physics D*, *15*(11), 1753–1935.

Corter, J. E., Esche, S. K., Chassapis, C., Ma, J. & Nickerson, J. V. (2011). Process and learning outcomes from remotely-operated, simula-

ted, and hands-on student laboratories. *Computers & Education*, *57*(3), 2054–2067.

Dane, E. & Pratt, M. G. (2007). Exploring Intuition and its Role in Managerial Decision Making. *Academy of Management Review*, *32*(1), 33–54.

de Jong, T., Linn, M. C. & Zacharia, Z. C. (2013). Physical and Virtual Laboratories in Science and Engineering Education. *Science*, *340*(6130), 305–308.

de Jong, T., Martin, E., Zamarro, J.-M., Esquembre, F., Swaak, J. & van Joolingen, W. R. (1999). The Integration of Computer Simulation and Learning Support: An Example from the Physics Domain of Collisions. *Journal of Research in Science Teaching*, *36*(5), 597–615.

de Jong, T. & van Joolingen, W. R. (1998). Scientific Discovery Learning with Computer Simulations of Conceptual Domains. *Review of Educational Research*, *68*(2), 179–201.

DeBoer, G. E. (2000). Scientific literacy: Another look at its historical and contemporary meanings and its relationship to science education reform. *Journal of Research in Science Teaching*, *37*(6), 582–601.

Deci, E. L. & Ryan, R. M. (1993). Die Selbstbestimmungstheorie der Motivation und ihre Bedeutung für die Pädagogik. *Zeitschrift für Pädagogik*, *39*, 224–238.

Decker, R. (2012). *Das Pendel - Physikalische Grundlagen und Schülervorstellungen* (Bachelorarbeit, Ruhr-Universität Bochum).

Department for Education. (2015). National curriculum in England: science programmes of study. Zugriff 27. März 2017, unter https://www.gov.uk/government/publications/national-curriculum-in-england-science-programmes-of-study/national-curriculum-in-england-science-programmes-of-study

Dittmer, A. & Gebhard, U. (2012). Stichwort Bewertungskompetenz: Ethik im naturwissenschaftlichen Unterricht aus sozialintuitionistischer Perspektive. *Zeitschrift für Didaktik der Naturwissenschaften*, *18*(2012), 81–98.

Dole, J. A. & Sinatra, G. M. (1998). Reconceptalizing change in the cognitive construction of knowledge. *Educational Psychologist*, *33*(2), 109–128.

Döring, N. & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften*. Springer-Lehrbuch. Berlin, Heidelberg: Springer.

Driver, R., Newton, P. & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. *Science Education*, *84*(3), 287–312.

Duit, R. (1986). Wärmevorstellungen. *Naturwissenschaften im Unterricht - Physik/Chemie*, *34*(13), 30–33.

Duit, R. & Treagust, D. (2003). Conceptual Change - A powerful framework for improving science teaching and learning. *International Journal of Science Education*, *25*, 671–688.

Duit, R., von Zelewski, H.-D. & Heyner, W. (1978). *IPN Curriculum Physik - Unterrichtseinheiten für das 7. und 8. Schuljahr. Didaktische Anleitungen: Energie - Arbeit - Leistung - Kraft*. Stuttgart: Klett.

Duschl, R. A. (2007). Quality Argumentation and Epistemic Criteria. In S. Erduran & M. P. Jiménez-Aleixandre (Hrsg.), *Argumentation in Science Education* (35, S. 159–175). Science & Technology Education Library. Dordrecht: Springer Netherlands.

Eagly, A. H. & Chaiken, S. (1993). *The psychology of attitudes*. Orlando, FL, US: Harcourt Brace Jovanovich College Publishers.

Eid, M., Gollwitzer, M. & Schmitt, M. (2013). *Statistik und Forschungsmethoden* (3. Auflage). Weinheim: Beltz.

Emden, M. (2011). *Prozessorientierte Leistungsmessung des naturwissenschaftlichen-experimentellen Arbeitens. Eine vergleichende Studie zu Diagnoseinstrumenten zu Beginn der Sekundarstufe I*. Berlin: Logos Verlag.

Enders, C. K. & Bandalos, D. L. (2001). The Relative Performance of Full Information Maximum Likelihood Estimation for Missing Data in Structural Equation Models. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*(3), 430–457.

Epstein, S. (1998). Cognitive-Experiential Self-Theory. In D. F. Barone, M. Hersen & V. B. V. Hasselt (Hrsg.), *Advanced Personality* (S. 211–238). New York, NY: Springer.

Epstein, S., Pacini, R., Denes-Raj, V. & Heier, H. (1996). Individual differences in intuitive–experiential and analytical–rational thinking styles. *Journal of Personality and Social Psychology*, *71*(2), 390–405.

Erduran, S., Simon, S. & Osborne, J. (2004). TAPping into argumentation: Developments in the application of Toulmin's Argument Pattern for studying science discourse. *Science Education*, *88*(6), 915–933.

Feinstein, A. R. & Cicchetti, D. V. (1990). High agreement but low kappa: The problems of two paradoxes. *Journal of Clinical Epidemiology*, *43*(6), 543–549.

Fensham, P. P. & Marton, P. F. (1992). What has happened to intuition in science education? *Research in Science Education*, *22*(1), 114–122.

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice*, *40*(5), 532–538.

Field, A., Miles, J. & Field, Z. (2012). *Discovering Statistics Using R*. London: SAGE.

Finkelstein, N. D., Adams, W. K., Keller, C. J., Kohl, P. B., Perkins, K. K., Podolefsky, N. S., ... LeMaster, R. (2005). When Lear-

ning about the Real World is Better Done Virtually: A Study of Substituting Computer Simulations for Laboratory Equipment. *Physical Review Special Topics - Physics Education Research*, *1*, 1–8.

Finney, S. J. & DiStefano, C. (2013). Nonnormal and categorical data in structural equation models. In G. R. Hancock & R. O. Mueller (Hrsg.), *Structural Equation Modeling: A second course*. Charlotte, NC: Information Age Publishing.

Fischbein, H. (1987). *Intuition in Science and Mathematics: An Educational Approach*. Dordrecht: Springer Netherlands.

Ford, M. J. (2008). Disciplinary authority and accountability in scientific practice and learning. *Science Education*, *92*(3), 404–423.

Ford, M. J. (2012). A Dialogic Account of Sense-Making in Scientific Argumentation and Reasoning. *Cognition and Instruction*, *30*(3), 207–245.

Free Software Foundation. (2011). GNU General Public License, Version 2. Zugriff 12. Oktober 2011, unter http://www.gnu.org/licenses/gpl-2.0.html

Frey, A., Taskinen, P., Schütte, K., Prenzel, M., Artelt, C., Baumert, J., ... Pekrun, R. (2009). *PISA 2006 Skalenhandbuch: Dokumentation der Erhebungsinstrumente*. Münster: Waxmann.

Friege, G. & Lind, G. (2004). Leistungsmessung im Leistungskurs. *Der mathematische und naturwissenschaftliche Unterricht*, *57*(5), 259–265.

Furnham, A. & Thorne, J. D. (2013). Need for Cognition. *Journal of Individual Differences*, *34*(4), 230–240.

Ganser, M. & Hammann, M. (2009). Hypothesen verändern können. Aufgaben zum Umgang mit unerwarteten Daten im Kontext historischer Experimente. *Praxis der Naturwissenschaften – Biologie in der Schule*, *58*(3), 39–43.

Garcia-Mila, M. & Andersen, C. (2007). Cognitive Foundations of Learning Argumentation. In M. P. Jiménez-Aleixandre & S. Erduran (Hrsg.), *Argumentation in Science Education. Perspectives from Classroom-Based Research* (S. 29–45). Dordrecht: Springer.

Gaspard, H. (2015). *Promoting Value Beliefs in Mathematics: A Multidimensional Perspective and the Role of Gender* (Dissertation, Universität Tübingen).

Gigerenzer, G. & Gaissmaier, W. (2011). Heuristic Decision Making. *Annual Review of Psychology*, *62*(1), 451–482.

Gignac, G. E. & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, *102*, 74–78.

Gorsky, P. & Finegold, M. (1994). The role of anomaly and of cognitive dissonance in restructuring students' concepts of force. *Instructional Science*, *22*(2), 75–90.

Gößling, J. (2010). *Selbständig entdeckendes Experimentieren - Lernwirksamkeit der Strategieanwendung* (Dissertation, Universität Duisburg-Essen).

Gott, R. & Duggan, S. (1996). Practical work: its role in the understanding of evidence in science. *International Journal of Science Education*, *18*(7), 791–806.

Gott, R. & Duggan, S. (2007). A framework for practical work in science and scientific literacy through argumentation. *Research in Science & Technological Education*, *25*(3), 271–291.

Gott, R., Duggan, S. & Roberts, R. (2014). Concepts of Evidence. Zugriff unter http : / / community . dur . ac . uk / rosalyn . roberts / Evidence/cofev.htm

Graham, J. W. & Coffman, D. L. (2012). SEM with Missing Data. In R. H. Hoyle (Hrsg.), *Handbook of Structural Equation Modeling* (S. 277–302). New York: Guilford Press.

Green, S. B. & Thompson, M. S. (2012). A Flexible Structural Equation Modeling Approach for Analyzing Means. In R. H. Hoyle (Hrsg.), *Handbook of Structural Equation Modeling* (S. 393–416). New York: Guilford Press.

Greeno, J. G. (1998). The situativity of knowing, learning, and research. *American Psychologist*, *53*(1), 5.

Grehn, J. (Hrsg.). (1991). *Metzler Physik*. Stuttgart: J.B. Metzler.

Gromadecki, U. (2009). *Argumente in physikalischen Kontexten - Welche Geltungsgründe halten Physikanfänger für überzeugend?* Berlin: Logos Verlag.

Guzzetti, B. J., Snyder, T. E., Glass, G. V. & Gamas, W. S. (1993). Promoting Conceptual Change in Science: A Comparative Meta-Analysis of Instructional Interventions from Reading Education and Science Education. *Reading Research Quarterly*, *28*(2), 116–59.

Gwet, K. L. (2010). *Handbook of Inter-Rater Reliability*. Gaithersburg, MD: Advanced Analytics, LLC.

Haddock, C. K., Rindskopf, D. & Shadish, W. R. (1998). Using Odds Ratios as Effect Sizes for Meta-Analysis of Dichotomous Data: A Primer on Methods and Issues. *Psychological Methods*, *3*(3), 339–353.

Hammann, M. (2004). Kompetenzentwicklungsmodelle. Merkmale und ihre Bedeutung - dargestellt anhand von Kompetenzen beim Experimentieren. *Der mathematische und naturwissenschaftliche Unterricht*, *54*(3), 196–203.

Hammann, M. (2007). Das Scientific Discovery as Dual Search Modell. In D. Krüger & H. Vogt (Hrsg.), *Theorien in der biologiedidaktischen Forschung* (S. 187–196). Berlin/Heidelberg: Springer Verlag.

Hammann, M. (2010). Experimentieren. In W. Ruppert & P. D. U. Spörhase (Hrsg.), *Fachmethodik: Biologie-Methodik: Handbuch für die Sekundarstufe I und II* (S. 87–91). Berlin: Cornelsen: Scriptor.

Hammann, M., Phan, T. H., Ehmer, M. & Bayrhuber, H. (2006). Fehlerfrei Experimentieren. *Der mathematische und naturwissenschaftliche Unterricht*, *59*(5), 292–299.

Hammer, D. & Berland, L. K. (2014). Confusing Claims for Data: A Critique of Common Practices for Presenting Qualitative Research on Learning. *Journal of the Learning Sciences*, *23*(1), 37–46.

Hancock, G. R. (2001). Effect size, power, and sample size determination for structured means modeling and mimic approaches to between-groups hypothesis testing of means on a single latent construct. *Psychometrika*, *66*(3), 373–388.

Hansen, H. (2015). *Fallacies*. In E. N. Zalta (Hrsg.), *The Stanford Encyclopedia of Philosophy* (Summer 2015 Edition). Metaphysics Research Lab, Stanford University. Zugriff 3. März 2017, unter https://plato.stanford.edu/archives/sum2015/entries/fallacies/

Hartig, J., Frey, A. & Jude, N. (2012). Validität. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 143–171). Berlin-Heidelberg: Springer.

Heinicke, S. (2012). *Aus Fehlern wird man klug: Eine genetisch-didaktische Rekonstruktion des "Messfehlers"*. Berlin: Logos.

Heitmann, P., Hecht, M., Schwanewedel, J. & Schipolowski, S. (2014). Students' Argumentative Writing Skills in Science and First-Language Education: Commonalities and differences. *International Journal of Science Education*, *36*(18), 3148–3170.

Hellwig, J. (2012). *Messunsicherheiten verstehen: Entwicklung eines normativen Sachstrukturmodells am Beispiel des Unterrichtsfaches Physik* (Dissertation, Ruhr-Universität Bochum).

Henson, R. K. & Roberts, J. K. (2006). Use of Exploratory Factor Analysis in Published Research Common Errors and Some Comment on Improved Practice. *Educational and Psychological Measurement*, *66*(3), 393–416.

Hevey, D., Thomas, K., Pertl, M., Maher, L., Craig, A. & Chuinneagain, S. N. (2012). Method Effects and the Need for Cognition Scale. *International Journal of Educational and Psychological Assessment*, *12*, 20–33.

Hidi, S. & Renninger, K. A. (2006). The Four-Phase Model of Interest Development. *Educational Psychologist*, *41*(2), 111–127.

Hodgkinson, G. P., Langan-Fox, J. & Sadler-Smith, E. (2008). Intuition: A fundamental bridging construct in the behavioural sciences. *British Journal of Psychology*, *99*(1), 1–27.

Hodson, D. (1993). Re-thinking Old Ways: Towards A More Critical Approach To Practical Work In School Science. *Studies in Science Education*, *22*(1), 85–142.

Hofstein, A. (2017). The Role of Laboratory in Science Teaching and Learning. In K. S. Taber & B. Akpan (Hrsg.), *Science Education: An International Course Companion* (S. 357–368). Rotterdam: Sense.

Hofstein, A. & Lunetta, V. N. (2004). The laboratory in science education: Foundations for the twenty-first century. *Science Education*, *88*(1), 28–54.

Hogarth, R. M. (2001). *Educating Intuition*. Chicago: University of Chicago Press.

Hosmer, D. W., Lemeshow, S. & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3. Aufl.). Hoboken, NJ: Wiley.

Höttecke, D. & Rieß, F. (2015). Naturwissenschaftliches Experimentieren im Lichte der jüngeren Wissenschaftsforschung – Auf der Suche nach einem authentischen Experimentbegriff der Fachdidaktik. *Zeitschrift für Didaktik der Naturwissenschaften*, *21*(1), 127–139.

Hoyle, R. H. (2012). *Handbook of Structural Equation Modeling*. Guilford Press.

Hu, L. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55.

Hug, B. & McNeill, K. L. (2008). Use of First-hand and Second-hand Data in Science: Does data type influence classroom conversations? *International Journal of Science Education*, *30*(13), 1725–1751.

Hulleman, C. S., Godes, O., Hendricks, B. L. & Harackiewicz, J. M. (2010). Enhancing interest and performance with a utility value intervention. *Journal of Educational Psychology*, *102*(4), 880–895.

Hulleman, C. S. & Harackiewicz, J. M. (2009). Promoting Interest and Performance in High School Science Classes. *Science*, *326*(5958), 1410–1412.

International Association for the Evaluation of Educational Achievement. (1997). *TIMSS science items: Release set for population 2 (seventh and eighth grades)*. Boston College: TIMSS & PIRLS International Study Center: Lynch School of Education. Zugriff unter http://timssandpirls.bc.edu/timss1995i/Items.html

International Association for the Evaluation of Educational Achievement. (2001). *TIMSS 1999: Science items. Released set for eighth grade*. Boston College: TIMSS & PIRLS International Study Center: Lynch School of Education. Zugriff unter http://timssandpirls.bc.edu/timss1999i/study.html

International Association for the Evaluation of Educational Achievement. (2007). *TIMSS 2003: Science items. Released set eighth grade*. Boston College: TIMSS & PIRLS International Study Center: Lynch School of Education. Zugriff unter http://timssandpirls.bc.edu/timss2003i/released.html

International Association for the Evaluation of Educational Achievement. (2009). *TIMSS 2007: User guide for the international database. Released items. Science – fourth grade*. Boston College: TIMSS & PIRLS International Study Center: Lynch School of Education.

Zugriff unter http://timssandpirls.bc.edu/TIMSS2007/idb_ug.html

Iordanou, K. & Constantinou, C. P. (2015). Supporting Use of Evidence in Argumentation Through Practice in Argumentation and Reflection in the Context of SOCRATES Learning Environment. *Science Education*, *99*(2), 282–311.

Irribarra, D. T. & Freund, R. (2014). *Wright Map: IRT item-person map with ConQuest integration*. Zugriff unter http://github.com/david-ti/wrightmap

Jamieson, S. (2004). Likert scales: how to (ab)use them. *Medical Education*, *38*, 1217–1218.

Jiménez-Aleixandre, M. P. (2007). Designing Argumentation Learning Environments. In S. Erduran & M. P. Jiménez-Aleixandre (Hrsg.), *Argumentation in Science Education* (S. 91–117). Science & Technology Education Library. Dordrecht: Springer Netherlands.

Jiménez-Aleixandre, M. P. & Erduran, S. (2007). Argumentation in Science Education: An Overview. In M. P. Jiménez-Aleixandre & S. Erduran (Hrsg.), *Argumentation in Science Education. Perspectives from Classroom-Based Research* (S. 3–29). Dordrecht: Springer.

Josset, T., Perez, A. & Sudarsky, D. (2017). Dark Energy from Violation of Energy Conservation. *Physical Review Letters*, *118*(2), 021102.

Kahneman, D. & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*(4), 237–251.

Kanari, Z. & Millar, R. (2004). Reasoning from data: How students collect and interpret data in science investigations. *Journal of Research in Science Teaching*, *41*(7), 748–769.

Kang, H., Scharmann, L. C., Kang, S. & Noh, T. (2010). Cognitive Conflict and Situational Interest as Factors Influencing Conceptual Change. *International Journal of Environmental and Science Education*, *5*(4), 383–405.

Katchevich, D., Hofstein, A. & Mamlok-Naaman, R. (2013). Argumentation in the Chemistry Laboratory: Inquiry and Confirmatory Experiments. *Research in Science Education*, *43*(1), 317–345.

Kauertz, A. (2008). *Schwierigkeitserzeugende Merkmale physikalischer Leistungstestaufgaben*. Berlin: Logos Verlag.

Kebritchi, M., Hirumi, A. & Bai, H. (2010). The effects of modern mathematics computer games on mathematics achievement and class motivation. *Computers & Education*, *55*(2), 427–443.

Keller, J., Bohner, G. & Erb, H.-P. (2000). Intuitive und heuristische Urteilsbildung - verschiedene Prozesse? *Zeitschrift für Sozialpsychologie*, *31*(2), 87–101.

Kelly, G. J. (2008). Inquiry, Activity, and Epistemic Practice. In R. A. Duschl & R. Grandy (Hrsg.), *Teaching Scientific Inquiry: Recommendations for Research and Implementation* (S. 99–117). Rotterdam: Sense Publishers.

Kelly, G. J., Druker, S. & Chen, C. (1998). Students' reasoning about electricity: combining performance assessments with argumentation analysis. *International Journal of Science Education*, *20*(7), 849–871.

Kelly, G. J., Regev, J. & Prothero, W. (2007). Analysis of Lines of Reasoning in Written Argumentation. In S. Erduran & M. P. Jiménez-Aleixandre (Hrsg.), *Argumentation in Science Education. Perspectives from Classroom-Based Research* (S. 3–29). Dordrecht: Springer.

Kelly, G. J. & Takao, A. (2002). Epistemic levels in argument: An analysis of university oceanography students' use of evidence in writing. *Science Education*, *86*(3), 314–342.

Kenny, D. (2014). Measuring Model Fit. Zugriff 14. Januar 2014, unter http://davidakenny.net/cm/fit.htm

Kiefer, T., Robitzsch, A. & Wu, M. (2016). TAM: Test Analysis Modules (Version 1.16-0). Zugriff 2. März 2016, unter https://cran.r-project.org/web/packages/TAM/index.html

Kim, H. & Song, J. (2006). The Features of Peer Argumentation in Middle School Students' Scientific Inquiry. *Research in Science Education*, *36*(3), 211–233.

Kind, P. M., Kind, V., Hofstein, A. & Wilson, J. (2011). Peer Argumentation in the School Science Laboratory—Exploring effects of task features. *International Journal of Science Education*, *33*(18), 2527–2558.

Kirkup, L. & Frenkel, B. (2006). *An Introduction to Uncertainty in Measurement: Using the GUM*. Guide to the Expression of Uncertainty in Measurement. Cambridge: Cambridge University Press.

Kirschner, S., Borowski, A. & Fischer, H. (2011). ProwiN-Test zum Fachwissen von Physiklehrkräften. In A. Borowski, H. Fischer, M. Jüttner, S. Kirschner, B. J. Neuhaus, E. Sumfleth, ... S. Witner (Hrsg.), *ProwiN-Testinstrumente*. Essen: Universität Duisburg-Essen.

Klahr, D. (2000). *Exploring Science: The Cognition and Development of Discovery Processes*. Cambridge: MIT Press.

Klahr, D. & Dunbar, K. (1988). Dual Space Search During Scientific Reasoning. *Cognitive Science*, *12*(1), 1–48.

Klahr, D., Triona, L. M. & Williams, C. (2007). Hands on what? The relative effectiveness of physical versus virtual materials in an engineering design project by middle school children. *Journal of Research in Science Teaching*, *44*(1), 183–203.

Klayman, J. & Ha, Y.-W. (1987). Confirmation, Disconfirmation, and Information in Hypothesis Testing. *Psychological Review*, *94*(2), 211–228.

Kline, R. B. (2011). *Principles and Practice of Structural Equation Modeling*. New York: Guilford Press.

Kline, R. B. (2013). *Beyond significance testing: Statistics reform in the behavioral sciences* (2. Aufl.). Washington: American Psychological Association.

Kline, R. B. (2016). *Principles and Practice of Structural Equation Modeling* (4. Aufl.). Methodology in the Social Sciences. The Guilford Press.

Knapp, T. R. (1990). Treating Ordinal Scales as Interval Scales: An Attempt To Resolve the Controversy. *Nurs Res*, *39*(2), 121–123.

Knogler, M., Harackiewicz, J. M., Gegenfurtner, A. & Lewalter, D. (2015). How situational is situational interest? Investigating the longitudinal structure of situational interest. *Contemporary Educational Psychology*, *43*, 39–50.

Knogler, M. & Lewalter, D. (2014). Design-Based Research im naturwissenschaftlichen Unterricht. Das motivationsfördernde Potenzial situierter Lernumgebungen im Fokus. *Psychologie in Erziehung und Unterricht*, *61*(1), 2–14.

Kolstø, S. D. & Ratcliffe, M. (2007). Social Aspects of Argumentation. In M. P. Jiménez-Aleixandre & S. Erduran (Hrsg.), *Argumentation in Science Education. Perspectives from Classroom-Based Research* (S. 117–136). Dordrecht: Springer.

Koslowski, B. (1996). *Theory and Evidence: The Development of Scientific Reasoning*. MIT Press.

Krapp, A. (2001). Interesse. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (S. 286–294). Weinheim: Beltz.

Krapp, A. (2002). An educational-psychological theory of interest and its relation to self-determination theory. In E. L. Deci & R. M. Ryan (Hrsg.), *The handbook of self-determination research* (S. 405–427). Rochester: University of Rochester Press.

Krapp, A. & Prenzel, M. (2011). Research on Interest in Science: Theories, methods, and findings. *International Journal of Science Education*, *33*(1), 27–50.

Krüger, D. & Riemeier, T. (2014). Die qualitative Inhaltsanalyse - eine Methode zur Auswertung von Interviews. In D. Krüger, I. Parchmann & H. Schecker (Hrsg.), *Methoden in der naturwissenschaftsdidaktischen Forschung*. Berlin, Heidelberg: Springer Berlin Heidelberg. Zugriff 8. März 2017, unter http://link.springer.com/10.1007/978-3-642-37827-0

Kuhn, D. (1993). Science as argument: Implications for teaching and learning scientific thinking. *Science Education*, *77*(3), 319–337.

Kuhn, D. & Udell, W. (2003). The development of argument skills. *Child Development*, *74*(5), 1245–1260.

Künsting, J., Thillmann, H., Wirth, J., Fischer, H. E. & Leutner, D. (2008). Strategisches Experimentieren im naturwissenschaftlichen Unterricht. *Psychologie in Erziehung und Unterricht*, *55*, 1–15.

Künsting, J., Wirth, J. & Paas, F. (2011). The goal specificity effect on strategy use and instructional efficiency during computer-based scientific discovery learning. *Computers & Education*, *56*(3), 668–679.

Landerman, L. R., Mustillo, S. A. & Land, K. C. (2011). Modeling repeated Measures of Dichotomous Data. *Social science research*, *40*(5), 1456–1464.

Lau, R. (2013). *Argumentationen bei nicht-hypothesenkonformen experimentellen Daten an einem Beispiel aus der Wärmelehre*. Masterarbeit. Humboldt-Universität zu Berlin.

Lave, J. & Wenger, E. (1991). *Situated Learning: Legitimate Peripheral Participation*. Cambridge University Press.

Leach, J. & Scott, P. (2003). Individual and Sociocultural Views of Learning in Science Education. *Science & Education*, *12*(1), 91–113.

Lederman, J. S., Lederman, N. G., Bartos, S. A., Bartels, S. L., Meyer, A. A. & Schwartz, R. S. (2014). Meaningful assessment of learners' understandings about scientific inquiry—The views about scientific inquiry (VASI) questionnaire. *Journal of Research in Science Teaching*, *51*(1), 65–83.

Lee, G. & Byun, T. (2011). An Explanation for the Difficulty of Leading Conceptual Change Using a Counterintuitive Demonstration: The Relationship Between Cognitive Conflict and Responses. *Research in Science Education*, *42*(5), 943–965.

Lee, T., Cai, L., MacCallum, R. C. & Hoyle, R. H. (2012). Power Analysis for Tests of Structural Equation Models. In *Handbook of Structural Equation Modeling* (S. 181–194). New York: Guilford Press.

Lehrer, R., Kim, M.-j. & Schauble, L. (2007). Supporting the Development of Conceptions of Statistics by Engaging Students in Measuring and Modeling Variability. *International Journal of Computers for Mathematical Learning*, *12*(3), 195–216.

Lemke, J. L. (1990). *Talking Science: Language, Learning, and Values*. Westport, CT: Ablex Publishing Corporation.

Lewalter, D. & Knogler, M. (2014). A Questionnaire to Assess Situational Interest - Theoretical Considerations and Findings. Annual Meeting of the American Educational Research Association. Philadelphia, PA.

Lienert, G. A. & Raatz, U. (1998). *Testaufbau und Testanalyse*. Weinheim: Beltz, Psychologie Verl.-Union.

Lin, J.-Y. (2007). Responses to anomalous data obtained from repeatable experiments in the laboratory. *Journal of Research in Science Teaching*, *44*(3), 506–528.

Lipschutz, S., Spiegel, M. R. & Liu, J. (2009). *Schaum's Outline of Mathematical Handbook of Formulas and Tables: 2,400 Formulas + Tables* (3. Aufl.). New York: McGraw-Hill.

Little, T. D., Rhemtulla, M., Gibson, K. & Schoemann, A. M. (2013). Why the items versus parcels controversy needn't be one. *Psychological Methods*, *18*(3), 285–300.

Lord, K. R. & Putrevu, S. (2006). Exploring the dimensionality of the need for cognition scale. *Psychology & Marketing*, *23*(1), 11–34.

Lüders, K. & von Oppen, G. (2012). *Klassische Physik - Mechanik und Wärme*. Berlin: DeGruyter.

Lüdtke, O., Robitzsch, A., Trautwein, U. & Köller, O. (2007). Umgang mit fehlenden Werten in der psychologischen Forschung. *Psychologische Rundschau*, *58*(2), 103–117.

Ludwig, T. (2011). *Begründungen und Überzeugungen hinsichtlich der Richtigkeit selbst aufgestellter Hypothesen in Real- und Simulationsexperimenten* (Masterarbeit, Ruhr-Universität Bochum).

Lunetta, V. N., Hofstein, A. & Clough, M. P. (2007). Learning and Teaching in the School Science Laboratory: An Analysis of Research, Theory, and Practice. In S. K. Abell & N. G. Lederman (Hrsg.), *Handbook of Research on Science Education* (S. 393–442). London: Lawrence Erlbaum.

MacCallum, R. C., Browne, M. W. & Cai, L. (2006). Testing differences between nested covariance structure models: Power analysis and null hypotheses. *Psychological Methods*, *11*(1), 19–35.

MacCallum, R. C., Browne, M. W. & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*(2), 130–149.

Maio, G. & Haddock, G. (2009). *The Psychology of Attitudes and Attitude Change*. SAGE.

Manz, E. (2014). Representing Student Argumentation as Functionally Emergent From Scientific Activity. *Review of Educational Research*, 0034654314558490.

Marcus-Roberts, H. M. & Roberts, F. S. (1987). Meaningless Statistics. *Journal of Educational Statistics*, *12*(4), 383–394.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, *57*(3), 519–530.

Masnick, A. M., Klahr, D. & Knowles, E. R. (2017). Data-Driven Belief Revision in Children and Adults. *Journal of Cognition and Development*, *18*(1), 87–109.

Mason, L. (2001). Responses to anomalous data on controversial topics and theory change. *Learning and Instruction*, *11*(6), 453–483.

Mayer, J., Keiner, K., Ziemek, H.-P. & Klee, R. (2003). Naturwissenschaftliche Problemlösekompetenz im Biologieunterricht. In A. Bauer & H. Bayrhuber (Hrsg.), *Entwicklung von Wissen und Kompetenzen im Biologieunterricht* (S. 21–24). Kiel: IPN.

McComas, W. F. (2004). Keys to Teaching the Nature of Science. *The Science Teacher*, *71*(9), 24–27.

McKelvey, R. D. & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *The Journal of Mathematical Sociology*, *4*(1), 103–120.

McNeill, K. L. & Krajcik, J. (2007). Middle School Students' Use of Appropriate and Inappropriate Evidence in Writing Scientific Explanations. In M. Lovett & P. Shah (Hrsg.), *Thinking with Data* (S. 233–265). Taylor & Francis Group.

Meade, A. W., Johnson, E. C. & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, *93*(3), 568–592.

Messick, S. (1989). Validity. In R. L. Linn (Hrsg.), *Educational measurement* (S. 13–103). Washington, D.C.: American Council on Education and National Council on Measurement in Education.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749.

Meyer, P. (2010). *Understanding Measurement: Reliability*. Oxford: Oxford University Press.

Miller, G. A. & Chapman, J. P. (2001). Misunderstanding analysis of covariance. *Journal of Abnormal Psychology*, *110*(1), 40–48.

Ministerium für Schule und Weiterbildung, Wissenschaft und Forschung des Landes Nordrhein-Westfalen (Hrsg.). (1999). *Lehrplan Physik*. Frechen: Ritterbach Verlag.

Ministerium für Schule und Weiterbildung, Wissenschaft und Forschung des Landes Nordrhein-Westfalen (Hrsg.). (2008). *Kernlehrplan für das Gymnasium - Sekundarstufe I in Nordrhein-Westfalen - Physik*. Frechen: Ritterbach Verlag.

Mitchell, M. (1993). Situational interest: Its multifaceted structure in the secondary school mathematics classroom. *Journal of Educational Psychology*, *85*(3), 424–436.

Moosbrugger, H. & Kevala, A. (2008). *Testtheorie und Fragebogenkonstruktion*. Berlin/Heidelberg: Springer Verlag.

Morse, J. M. & Niehaus, L. (2009). *Mixed method design: principles and procedures*. Walnut Creek, CA: Left Coast Press.

Muckenfuß, H. (1995). *Lernen im sinnstiftenden Kontext: Entwurf einer zeitgemäßen Didaktik des Physikunterrichts*. Berlin: Cornelsen Verlag.

Munier, V., Merle, H. & Brehelin, D. (2013). Teaching Scientific Measurement and Uncertainty in Elementary School. *International Journal of Science Education*, *35*(16), 2752–2783.

Murphy, P. K. & Mason, L. (2006). Changing Knowledge and Beliefs. In P. H. Winne & P. A. Alexander (Hrsg.), *Handbook of Educational Psychology* (2. Aufl., S. 305–324). Mahwah, NJ: Lawrence Erlbaum.

Muthén, B. O. (1998–2004). *Mplus Technical Apendices* (7. Aufl.). Los Angeles, CA: Muthén & Muthén.

Muthén, B. O. & Muthén, L. K. (1998–2015). *Mplus User's Guide* (7. Aufl.). Los Angeles, CA: Muthén & Muthén.

Naumann, J. (2004). *Unterschiede zwischen kognitionsbasierten und affektbasierten Einstellungen* (Dissertation, Universität Köln).

Nawrath, D., Maiseyenka, V. & Schecker, H. (2011). Experimentelle Kompetenz – Ein Modell für die Unterrichtspraxis. *Praxis der Naturwissenschaften - Physik in der Schule*, *60*(6), 42–49.

Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology*, *15*(3), 263–280.

Nelson, R. A. & Olsson, M. G. (1986). The pendulum - Rich physics from a simple system. *American Journal of Physics*, *54*(2), 112–121.

Neugebauer, C. (2006). *Lernen mit Simulationen und der Einfluss auf das Problemlösen in der Physik*. Berlin: Logos.

Neumann, K. (2014). Rasch-Analyse naturwissenschaftsbezogener Leistungstests. In D. Krüger, I. Parchmann & H. Schecker (Hrsg.), *Methoden in der naturwissenschaftsdidaktischen Forschung*. Berlin Heidelberg: Springer-Verlag.

NGSS Lead States. (2013). Next Generation Science Standards: For States, By States. Zugriff 27. Februar 2015, unter nextgenscience.org

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220.

Njoo, M. & de Jong, T. (1993). Exploratory Learning with a Computer Simulation for Control Theory: Learning Processes and Instructional Support. *Journal of Research in Science Teaching*, *30*(5), 821–844.

Nobel Media AB. (2014). *Transcript of the telephone interview with Saul Perlmutter following the announcement of the 2011 Nobel Prize in Physics*. Nobelprize.org. Zugriff 5. November 2011, unter http://www.nobelprize.org/nobel_prizes/physics/laureates/2011/perlmutter-telephone.html

Nolting, W. (2011). *Grundkurs Theoretische Physik 1: Klassische Mechanik* (2. Aufl.). Heidelberg: Springer.

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education: Theory and Practice*, *15*(5), 625–632.

Olivier, J., May, W. L. & Bell, M. L. (2016). Relative effect sizes for measures of risk. *Communications in Statistics - Theory and Methods*, 1–8.

Osborne, J. (2010). Arguing to Learn in Science: The Role of Collaborative, Critical Discourse. *Science*, *328*(5977), 463–466.

Osborne, J. (2013). The 21st century challenge for science education: Assessing scientific reasoning. *Thinking Skills and Creativity*, *10*, 265–279.

Osborne, J., Donovan, B. M., Henderson, J. B., MacPherson, A. C. & Wild, A. (2016). *Arguing From Evidence in Middle School Science: 24 Activities for Productive Talk and Deeper Learning*. Thousand Oaks, CA: Corwin.

Osborne, J., Erduran, S. & Simon, S. (2004). Enhancing the quality of argumentation in school science. *Journal of Research in Science Teaching*, *41*(10), 994–1020.

Osborne, J. & Patterson, A. (2011). Scientific argument and explanation: A necessary distinction? *Science Education*, *95*(4), 627–638.

Pacini, R. & Epstein, S. (1999). The relation of rational and experiential information processing styles to personality, basic beliefs, and the ratio-bias phenomenon. *Journal of Personality and Social Psychology*, *76*(6), 972.

Pechtl, H. (2009). *Anmerkungen zur Operationalisierung und Messung des Konstrukts need for cognition*. Wirtschaftswissenschaftliche Diskussionspapiere. Ernst-Moritz-Arndt-Universität Greifswald. Zugriff 10. Februar 2014, unter http : / / www . rsf . uni - greifswald . de / fileadmin / mediapool / Fakult _ t / Lenz / Diskussionspapiere/05-2009.pdf

Peebles, P. J. E. & Ratra, B. (2003). The cosmological constant and dark energy. *Reviews of Modern Physics*, *75*(2), 559–606.

Petty, R. E. & Cacioppo, J. T. (1984). Source Factors and the Elaboration Likelihood Model of Persuasion. *Advances in Consumer Research*, *11*, 668–672.

Petty, R. E. & Cacioppo, J. T. (1986). The Elaboration Likelihood Model of Persuasion. In L. Berkowitz (Hrsg.), *Advances in Experimental Social Psychology* (Bd. 19, S. 123–205). Orlando, Florida: Academic Press.

Petty, R. E. & Cacioppo, J. T. (1996). *Attitudes And Persuasion: Classic And Contemporary Approaches*. Oxford: Westview Press.

Petty, R. E., Cacioppo, J. T. & Goldman, R. (1981). Personal involvement as a determinant of argument-based persuasion. *Journal of Personality and Social Psychology*, *41*(5), 847–855.

Pfeiler, S. & Priemer, B. (2017). *Der Umgang mit Daten aus erster und zweiter Hand im Unterricht*. Jahrestagung des FV Didaktik der Physik der Deutschen Physikalischen Gesellschaft. Dresden.

PhET Interactive Simulations. (2011). Pendulum Lab. University of Colorado, Boulder. Zugriff 12. Oktober 2011, unter http://phet.colorado.edu/

Pintrich, P. R., Marx, R. W. & Boyle, R. A. (1993). Beyond Cold Conceptual Change: The Role of Motivational Beliefs and Classroom Contextual Factors in the Process of Conceptual Change. *Review of Educational Research*, *63*(2), 167–199.

Pornprasertmanit, S., Miller, P., Schoemann, A., Rosseel, Y., Quick, C., Garnier-Villarreal, M., … Chesnut, S. (2014). semTools: Useful tools for structural equation modeling (Version 0.4-6). Zugriff

24. Juni 2015, unter http://cran.r-project.org/web/packages/semTools/index.html

Posner, G. J., Strike, K. A., Hewson, P. & Gertzog, W. A. (1982). Accomodation of a scientific conception: Toward a theory of conceptual change. *Science Education*, *66*(2), 211–227.

Pospeschill, M. (2010). *Testtheorie, Testkonstruktion, Testevaluation*. München: Reinhardt.

Prechtl, P. (2016). *Philosophie*. Berlin: Springer-Verlag.

Pui-Wa, L. & Qiong, W. (2012). Estimation in Structural Equation Modeling. In R. H. Hoyle (Hrsg.), *Handbook of Structural Equation Modeling* (S. 164–180). New York: Guilford Press.

R Core Team. (2014). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Zugriff unter http://www.R-project.org/

Renken, M. D. & Nunez, N. (2013). Computer simulations and clear observations do not guarantee conceptual understanding. *Learning and Instruction*, *23*, 10–23.

Renninger, K. A. & Hidi, S. (2011). Revisiting the Conceptualization, Measurement, and Generation of Interest. *Educational Psychologist*, *46*(3), 168–184.

Revelle, W. (2015). Psych: Procedures for Psychological, Psychometric, and Personality Research (Version 1.5.4). Zugriff 24. Juni 2015, unter http://cran.r-project.org/web/packages/psych/index.html

Rhemtulla, M., Brosseau-Liard, P. É. & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354–373.

Richman, W. L., Kiesler, S., Weisband, S. & Drasgow, F. (1999). A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews. *Journal of Applied Psychology*, *84*(5), 754–775.

Riemeier, T., von Aufschnaiter, C., Fleischhauer, J. & Rogge, C. (2012). Argumentationen von Schülern prozessbasiert analysieren: Ansatz, Vorgehen, Befunde und Implikationen. *Zeitschrift für Didaktik der Naturwissenschaften*, *18*, 181–200.

Rohrmann, B. (1978). Empirische Studien zur Entwicklung von Antwortskalen für die sozialwissenschaftliche Forschung. *Zeitschrift für Sozialpsychologie*, *9*(3), 222–245.

Rosseel, Y. (5 2012). Lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36.

Rost, J. (2003). Zeitgeist und Moden empirischer Analysemethoden. *Forum Qualitative Sozialforschung*, *4*(2).

Rost, J. (2004). *Lehrbuch Testtheorie - Testkonstruktion*. Bern: Huber.

Roth, W.-M. (2004). "Tappen im Dunkeln": der Umgang mit Unsicherheiten und Unwägbarkeiten während des Forschungsprozesses. *Zeitschrift für qualitative Bildungs-, Beratungs- und Sozialforschung*, *5*(2), 155–178.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Rutten, N., van Joolingen, W. R. & van der Veen, J. T. (2012). The learning effects of computer simulations in science education. *Computers & Education*, *58*(1), 136–153.

Ryu, S. & Sandoval, W. A. (2012). Improvements to elementary children's epistemic understanding from sustained argumentation. *Science Education*, *96*(3), 488–526.

Sadler, T. D. (2009). Situated learning in science education: socioscientific issues as contexts for practice. *Studies in Science Education*, *45*(1), 1–42.

Sampson, V. & Clark, D. B. (2008). Assessment of the ways students generate arguments in science education: Current perspectives and recommendations for future directions. *Science Education*, *92*(3), 447–472.

Sampson, V., Enderle, P. & Grooms, J. (2013). Development and Initial Validation of the Beliefs About Reformed Science Teaching and Learning (BARSTL) Questionnaire. *School Science and Mathematics*, *113*(1), 3–15.

Sander, F., Schecker, H. & Niedderer, H. (2003). Computer Tools in the Lab - Effects Linking Theory and Experiment. In D. Psillos & H. Niedderer (Hrsg.), *Teaching and Learning in the Science Laboratory* (16, S. 219–230). Science & Technology Education Library. Dordrecht: Springer Netherlands.

Sandoval, W. A. (2003). Conceptual and Epistemic Aspects of Students' Scientific Explanations. *Journal of the Learning Sciences*, *12*(1), 5–51.

Sandoval, W. A. (2005). Understanding students' practical epistemologies and their influence on learning through inquiry. *Science Education*, *89*(4), 634–656.

Sandoval, W. A. (2012). Situating epistemological development. In *Volume 1: Full Papers* (S. 347–354). International Society of the Learning Sciences. Sydney.

Sandoval, W. A. (2014). Conjecture Mapping: An Approach to Systematic Educational Design Research. *Journal of the Learning Sciences*, *23*(1), 18–36.

Sandoval, W. A. & Millwood, K. A. (2005). The Quality of Students' Use of Evidence in Written Scientific Explanations. *Cognition and Instruction*, *23*(1), 23–55.

Sandoval, W. A. & Millwood, K. A. (2007). What Can Argumentation Tell Us About Epistemology? In S. Erduran & M. P. Jiménez-

Aleixandre (Hrsg.), *Argumentation in Science Education* (35, S. 71–88). Dordrecht: Springer.

Sandoval, W. A. & Reiser, B. J. (2004). Explanation-driven inquiry: Integrating conceptual and epistemic scaffolds for scientific inquiry. *Science Education*, *88*(3), 345–372.

Satorra, A. & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*(4), 507–514.

Scheck, F. (2007). *Theoretische Physik 1: Mechanik* (8. Aufl.). Berlin/Heidelberg: Springer.

Schecker, H., Neumann, K., Theyßen, H., Eickhorst, B. & Dickmann, M. (2016). Stufen experimenteller Kompetenz. *Zeitschrift für Didaktik der Naturwissenschaften*, *22*(1), 197–213.

Schermelleh-Engel, K., Moosbrugger, H. & Müller, H. (2003). Evaluating the Fit of Structural Equation Models: Tests of Significance and Descriptive Goodness-of-Fit Measures. *Methods of Psychological Research*.

Schiepe-Tiska, A., Schöps, K., Rönnebeck, S., Köller, O. & Prenzel, M. (2012). Naturwissenschaftliche Kompetenz in PISA 2012: Ergebnisse und Herausforderungen. In M. Prenzel, E. Sälzer, E. Klieme & O. Köller (Hrsg.), *PISA 2012. Fortschritte und Herausforderungen in Deutschland* (S. 189–219). Münster: Waxmann.

Schmidkunz, H. & Lindemann, H. (2003). *Das Forschend-entwickelnde Unterrichtsverfahren: Problemlösen im naturwissenschaftlichen Unterricht*. Hohenwarsleben: Westarp Wissenschaften.

Schmiemann, P. & Lücken, M. (2013). Validität - Misst mein Test, was er soll? In D. Krüger, I. Parchmann & H. Schecker (Hrsg.), *Methoden in der naturwissenschaftsdidaktischen Forschung*. Berlin/Heidelberg: Springer.

Schreiber, N. (2012). *Diagnostik Experimenteller Kompetenz: Validierung technologiegestützter Testverfahren im Rahmen eines Kompetenzstrukturmodells*. Berlin: Logos.

Schreiber, N., Theyßen, H. & Schecker, H. (2009). Experimentelle Kompetenz messen?! *PhyDid A - Physik und Didaktik in Schule und Hochschule*, *3*(8), 92–101.

Schulz, J. & Priemer, B. (2016). Development of an Assessment Tool to Probe Students' Understanding of Measurement Uncertainty. In *2016 ASEE Mid-Atlantic Section Conference*.

Schwarz, B. B. & Asterhan, C. S. (2010). Argumentation and reasoning. In *International Handbook of Psychology in Education* (S. 137–176). Bingley: Emerald Group Publishing.

Schwarz, B. B., Neuman, Y., Gil, J. & Ilya, M. (2003). Construction of Collective and Individual Knowledge in Argumentative Activity. *Journal of the Learning Sciences*, *12*(2), 219–256.

Scranton, R., Connolly, A. J., Nichol, R. C., Stebbins, A., Szapudi, I., Eisenstein, D. J., … Xu, Y. (2003). Physical Evidence for Dark

Energy. arXiv: astro-ph/0307335. Zugriff 9. Mai 2017, unter http://arxiv.org/abs/astro-ph/0307335

Sedlmeier, P., Lovett, M. & Priti, S. (2007). Statistical Reasoning: Valid Intuitions Put to Use. In *Thinking with Data*. New York London: Taylor & Francis Group.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (Hrsg.). (2004). *Bildungsstandards im Fach Physik für den Mittleren Schulabschluss. [German science standards for physics in middle schools]*. München: Wolters Kluwer.

Senatsverwaltung für Bildung, Jugend und Sport. (2006). *Rahmenlehrplan für die Sekundarstufe I - Physik*. Berlin.

Serra, P., Cooray, A., Holz, D. E., Melchiorri, A., Pandolfi, S. & Sarkar, D. (2009). No evidence for dark energy dynamics from a global analysis of cosmological data. *Physical Review D*, *80*(12), 121302.

Shepardson, D. P. (1999). The role of anomalous data in restructuring fourth graders' frameworks for understanding electric circuits. *International Journal of Science Education*, *21*(1), 77–94.

Siegel, M. A. & Ranney, M. A. (2003). Developing the changes in attitude about the relevance of science (CARS) questionnaire and assessing two high school science classes. *Journal of Research in Science Teaching*, *40*(8), 757–775.

Sinatra, G. M. (2005). The Warming Trend in Conceptual Change Research: The Legacy of Paul R. Pintrich. *Educational Psychologist*, *40*(2), 107–115.

Sinatra, G. M. & Kardash, C. M. (2004). Teacher candidates' epistemological beliefs, dispositions, and views on teaching as persuasion. *Contemporary Educational Psychology*, *29*(4), 483–498.

Smetana, L. K. & Bell, R. L. (2012). Computer Simulations to Support Science Instruction and Learning: A critical review of the literature. *International Journal of Science Education*, *34*(9), 1337–1370.

Steinle, F. (1997). Entering New Fields: Exploratory Uses of Experimentation. *Philosophy of Science*, *64*, 64–74.

Temme, D. & Hildebrandt, L. (2008). *Gruppenvergleiche bei hypothetischen Konstrukten: die Prüfung der Übereinstimmung von Messmodellen mit der Strukturgleichungsmethodik* (Nr. 2008,042). SFB 649 discussion paper.

Tesch, M. & Duit, R. (2004). Experimentieren im Physikunterricht - Ergebnisse einer Videostudie. *Zeitschrift für Didaktik der Naturwissenschaften*, *10*, 51–69.

Thillmann, H. (2007). *Selbstreguliertes Lernen durch Experimentieren: Von der Erfassung zur Förderung*.

Thillmann, H., Gößling, J., Wirth, J. & Leutner, D. (2009). Strategisches Lernen mit interaktiven digitalen Medienverbünden. In R. Plötzner, T. Leuders & A. Wichert (Hrsg.), *Lernchance Computer:*

Strategien für das Lernen mit digitalen Medienverbünden (S. 89–108). Münster: Waxmann Verlag.

Tipler, P. A. & Mosca, G. (2006). *Physik für Wissenschaftler und Ingenieure* (2. Aufl.). München: Elsevier.

Toulmin, S. E. (1958). *The Uses of Argument*. New York, NY: Cambridge University Press.

Toulmin, S. E. (2003). *The Uses of Argument*. New York, NY: Cambridge University Press.

Triona, L. M. & Klahr, D. (2003). Point and Click or Grab and Heft: Comparing the Influence of Physical and Virtual Instructional Materials on Elementary School Students' Ability To Design Experiments. *Cognition and Instruction*, *21*(2), 149–73.

Tröbst, S., Hardy, I. & Möller, K. (2011). Die Förderung deduktiver Schlussfolgerungen bei Grundschulkindern in naturwissenschaftlichen Kontexten. *Unterrichtswissenschaft*, *39*(1), 7–20.

Urban, D. (1993). *Logit-Analyse: Statistische Verfahren zur Analyse von Modellen mit qualitativen Response-Variablen*. Stuttgart: Gustav Fischer.

Urban, D. & Mayerl, J. (2014). *Strukturgleichungsmodellierung: Ein Ratgeber für die Praxis*. Wiesbaden: Springer VS.

van Eemeren, F. H. & Grootendorst, R. (2004). *A Systematic Theory of Argumentation: The Pragma-dialectical Approach*. Cambridge: Cambridge University Press.

van Joolingen, W. R. & de Jong, T. (1997). An extended dual search space model of scientific discovery learning. *Instructional Science*, *25*(5), 307–346.

Vandenberg, R. J. & Lance, C. E. (2000). A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. *Organizational Research Methods*, *3*(1), 4–70.

Wächter, M. & Kauertz, A. (2013). Förderung argumentativer Fähigkeiten im Physikunterricht. In S. Bernholt (Hrsg.), *Inquiry-based Learning - Forschendes Lernen* (Bd. 33, S. 605–607). Gesellschaft für Didaktik der Chemie und Physik. Jahrestagung in Hannover 2012. Kiel: IPN.

Walton, D. (1990). What is reasoning? What is an argument? *The Journal of Philosophy*, *87*(8), 399–419.

Walton, D. (1995). *Arguments from Ignorance*. University Park, PA: Penn State University Press.

Walton, D. (2008). *Argumentation Schemes* (1. Aufl.). Cambridge ; New York: Cambridge University Press.

Walton, D. (2010). *The Place of Emotion in Argument*. Penn State Press.

Walton, D. (2016). *Argument Evaluation and Evidence*. Law, Governance and Technology Series. Cham: Springer. Zugriff 3. März 2017, unter http://link.springer.com/10.1007/978-3-319-19626-8

Warm, T. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, (54), 427–450.

Watson, J. R., Swain, J. R. L. & McRobbie, C. (2004). Students' discussions in practical scientific inquiries. *International Journal of Science Education*, *26*(1), 25–45.

Weijters, B. & Baumgartner, H. (2012). Misresponse to Reversed and Negated Items in Surveys: A Review. *Journal of Marketing Research*, *49*(5), 737–747.

Weinberger, A. & Fischer, F. (2006). A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers & Education*. Methodological Issues in Researching CSCL, *46*(1), 71–95.

Wertsch, J. V. (1991). A sociocultural approach to socially shared cognition. In L. B. Resnick, J. M. Levine & S. D. Teasley (Hrsg.), *Perspectives on socially shared cognition* (S. 85–100). Washington, DC: American Psychological Association.

West, S. G., Aaron, T. B. & Wu, W. (2012). Model Fit and Model Selection in Structural Equation Modeling. In R. H. Hoyle (Hrsg.), *Handbook of Structural Equation Modeling* (S. 209–231). New York: Guilford Press.

West, S. G., Curran, P. J. & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological methods*, *1*(1), 16–29.

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software*, *21*(12), 1–20.

Wickham, H. (2011). The Split-Apply-Combine Strategy for Data Analysis. *Journal of Statistical Software*, *40*(1), 1–29.

Wickham, H. & Chang, W. (2015). Ggplot2: An Implementation of the Grammar of Graphics (Version 1.0.1). Zugriff 24. Juni 2015, unter http://cran.r-project.org/web/packages/ggplot2/index.html

Wild, E. & Möller, J. (2014). *Pädagogische Psychologie*. Springer-Verlag.

Williamson, D. S., Bangdiwala, S. I., Marshall, S. W. & Waller, A. E. (1996). Repeated measures analysis of binary outcomes: applications to injury research. *Accident; Analysis and Prevention*, *28*(5), 571–579.

Wirth, J., Künsting, J. & Leutner, D. (2009). The impact of goal specificity and goal type on learning outcome and cognitive load. *Computers in Human Behaviour*, *25*(2), 299–305.

Wirtz, M. & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität*. Göttingen: Hogrefe.

Wolf, E. J., Harrington, K. M., Clark, S. L. & Miller, M. W. (2013). Sample Size Requirements for Structural Equation Models An Evaluation of Power, Bias, and Solution Propriety. *Educational and Psychological Measurement*, 1–22.

Zacharia, Z. C. (2007). Comparing and combining real and virtual experimentation: an effort to enhance students' conceptual understanding of electric circuits. *Journal of Computer Assisted Learning*, *23*(2), 120–132.

Zacharia, Z. C. & Constantinou, C. P. (2008). Comparing the influence of physical and virtual manipulatives in the context of the Physics by Inquiry curriculum: The case of undergraduate students' conceptual understanding of heat and temperature. *American Journal of Physics*, *76*(4), 425–430.

Zacharia, Z. C., Loizou, E. & Papaevripidou, M. (2012). Is physicality an important aspect of learning through science experimentation among kindergarten students? *Early Childhood Research Quarterly*, *27*(3), 447–457.

Zacharia, Z. C. & Olympiou, G. (2011). Physical versus virtual manipulative experimentation in physics learning. *Learning and Instruction*, *21*(3), 317–331.

Zacharia, Z. C., Olympiou, G. & Papaevripidou, M. (2008). Effects of Experimenting with Physical and Virtual Manipulatives on Students' Conceptual Understanding in Heat and Temperature. *Journal of Research in Science Teaching*, *45*(9), 1021–1035.

Zander, S. (2016). *Lehrerfortbildung zu Basismodellen und Zusammenhänge zum Fachwissen*. Studien zum Physik- und Chemielernen. Berlin: Logos Verlag.

Zander, S., Krabbe, H. & Fischer, H. (2012). Entwicklung eines Fachwissenstests zur Mechanik für die Sekundarstufe I. In S. Bernholt (Hrsg.), *Konzepte fachdidaktischer Strukturierung für den Unterricht* (Bd. 32, S. 176–178). Gesellschaft für Didaktik der Chemie und Physik. Jahrestagung in Oldenburg 2011. Berlin: LIT Verlag.

Zeidler, D. L. (1997). The central role of fallacious thinking in science education. *Science Education*, *81*(4), 483–496.

Zimmerman, C. (2007). The Development of Scientific Thinking Skills in Elementary and Middle School. *Developmental Review*, *27*(2), 172–223.

Zohar, A. & Nemet, F. (2002). Fostering students' knowledge and argumentation skills through dilemmas in human genetics. *Journal of Research in Science Teaching*, *39*(1), 35–62.

DANKSAGUNG

Bei der Anfertigung dieser Arbeit haben eine Reihe Menschen in meinem Umfeld eine bedeutende Rolle gespielt. An dieser Stelle möchte ich mich herzlich bedanken bei

...meinem Doktorvater Burkhard Priemer, der bereits zu Studienzeiten mein Interesse an physikdidaktischen Fragestellungen erkannt und gefördert hat und mir die Möglichkeit zur wissenschaftlichen Qualifikation bot. Dass ich diese recht plötzlich statt in Bochum nun in Berlin vorantreiben sollte, sorgte für ein spannendes Abenteuer, dem ich mich gerne stellte. Insbesondere für die stets offene Bürotür, die maximale Freiheit in allen Belangen und das daraus resultierende selbstständige und eigenverantwortliche Arbeiten möchte ich mich herzlich bedanken.

...Doris Lewalter, für die vielen produktiven Projekttreffen in München oder Berlin, die stets von spannenden Diskussionen geprägt waren und durch das Erleben verschiedener catch- und hold-Phasen längerfristig zu einem individuellen Interesse am Interesse auf meiner Seite geführt haben.

...Alexander Kauertz, für die methodischen Beratungen ganz zu Beginn des Projekts und das fortwährende Interesse an dieser Arbeit.

...meinen Kolleginnen und Kollegen in Bochum, insbesondere Julia Hellwig, Stefan Kirchner und Irene Neumann, die mir den Einstieg ins Doktoranden-Dasein durch zahlreiche Ratschläge in den ersten Wochen erleichterten.

...den „alteingessessen“ Kolleginnen und Kollegen in Berlin, Franz Boczianowski, Marc Müller, Nico Westphal, Christina Lazai, Wiebke Musold, Patrick Meinhold und Frau Handtke, die mich herzlich aufgenommen haben, sowie bei meinen „neuen“ Kollegen Johannes Schulz, Steffen Wagner, Stephan Pfeiler, Daniel Zechlin, und Ulrike Gromadecki-Thiele. Vielen Dank für die intensiven Diskussionen im Rahmen von Journalclubs, Probevorträgen und in Kaffeepausen. Ich bin davon (auf der zentralen Route) überzeugt, dass ihr mit euren steten kritischen Nachfragen wesentlich dazu beigetragen habt, dass dieses Projekt gelingen konnte. Ein besonderer Dank geht an meinen langjährigen Bürokollegen Steffen für die Zweitbetreuung auf wissenschaftlicher und privater Ebene und die gesicherte Erkenntnis, dass sich die hartnäckigsten physikdidaktischen Probleme am schnellsten im Chagall lösen lassen.

...den Mitgliedern und Verantwortlichen der interdisziplinären MINT-Didaktik Graduiertenschule ProMINTion an der Humboldt-

Universität für die tollen Möglichkeiten zum Blick „über den Tellerrand“, zur interdisziplinären Diskussion und zur Weiterbildung in verschiedensten Bereichen. Besonders möchte ich mich bei Nora Butter und Kerstin Hasse für die Unterstützung bei der finanziellen Organisation des Auslandsaufenthaltes bedanken.

...bei Bodo Krause und Reinhard Beyer für die psychometrische Beratung ganz zu Beginn, sowie bei Ronny Scherer und Andrea Hildebrandt für die Hilfe bei den ersten konfirmatorischen Faktoranalysen.

...bei Simon Zander und Heiko Krabbe für die Möglichkeit, ihren Fachwissentest in meinen Studien einsetzen zu können.

...bei den Expertinnen und Experten, die den Testentwurf auf inhaltliche Validität geprüft haben.

...bei den vielen Lehrkräften, Schülerinnen und Schülern in Bochum und Berlin, die mich immer bereitwillig während der Entwicklungsstudien und der Hauptuntersuchung unterstützt haben. Insbesondere gebührt ein besonderer Dank den jeweiligen Fachleitern Physik der beteiligten Schulen, die durch ihr Engagement eine reibungslose und effiziente Datenerhebung ermöglicht haben.

...bei den Hilfskräften der Arbeitsgruppe, die mich so tatkräftig unterstützt haben: Marek, Jasmin, Alex, Franzi, Christina, Marcel, Jan, Laura, Kristoph, Christopher.

I am especially grateful to Bill Sandoval for giving me the opportunity to spend two very productive months at UCLA, pointing me gently towards a Situated Cognition perspective and making me rethink the fundamental assumptions of my work.

Ein besonderer Dank gebührt meinen Eltern und Großeltern, die mich auf diesem Weg immer bedingungslos unterstützt haben.

SELBSTSTÄNDIGKEITSERKLÄRUNG

Hiermit erkläre ich, die Dissertation selbstständig und nur unter Verwendung der angegebenen Hilfsmittel angefertigt zu haben.

Ich habe mich anderwärts nicht um einen Doktorgrad beworben und besitze einen entsprechenden Doktorgrad nicht.

Ich erkläre die Kenntnisnahme der dem Verfahren zugrunde liegenden Promotionsordnung der Mathematisch-Naturwissenschaftlichen Fakultät der Humboldt-Universität zu Berlin vom 27. Juni 2012.

Berlin, 15. Mai 2017

Tobias Ludwig